Cases and Materials on Privatization

Alexander Volokh
# Table of Contents

§ 0. Introduction.............................................................................................................................. 1
  § 0.1. Gillian E. Metzger, *Contemporary Examples of Privatization*, from *Privatization as Delegation*, 103 Colum. L. Rev. 1367, 1377–94 (2003)................................................................................................................................. 1

§ 1. The Non-Delegation Doctrine ................................................................................................ 24
  § 1.2. Panama Refining Co. v. Ryan, 293 U.S. 388 (1935)....................................................... 27
  § 1.3. A.L.A. Schechter Poultry Corp. v. United States, 295 U.S. 495 (1935)................................ 38
  § 1.5. Notes on the earlier federal cases ....................................................................................... 48
  § 1.8. Notes on the later federal cases ........................................................................................... 64
  § 1.9. Tex. Boll Weevil Eradication Found., Inc. v. Lewellen, 952 S.W.2d 454 (Tex. 1997) .................................................................................................................................................. 66
  § 1.10. General notes on the non-delegation doctrine ................................................................. 90

§ 2. The State Action Doctrine ..................................................................................................... 93
  § 2.1. The Civil Rights Cases, 109 U.S. 3 (1883)...................................................................... 93
  § 2.3. Notes on the early state-action cases ................................................................................... 102
  § 2.4. Flagg Brothers, Inc. v. Brooks, 436 U.S. 149 (1978).................................................... 105
  § 2.5. Lugar v. Edmonson Oil Co., 457 U.S. 922 (1982).......................................................... 113
  § 2.6. Notes on the “individual remedies” cases ......................................................................... 122
§ 2.10. Notes on the “quasi-governmental organization” cases .................. 148
§ 2.12. Notes on the conceptual difficulties of the state-action doctrine ....... 162

§ 3. State Action Applied: Private Police ................................................... 185

§ 4. Alternative Constitutional Proposals ................................................... 229
  § 4.4. Notes and questions...................................................................................... 241

§ 5. Why Are There Firms? ........................................................................ 271
  § 5.1. ADAM SMITH, THE WEALTH OF NATIONS (1776), Book I, Chapter I. Of the Division of Labour ................................................................. 271
  § 5.2. R.H. Coase, The Nature of the Firm, 4 ECONOMICA (n.s.) 386 (1937) ........ 275
  § 5.3. Armen A. Alchian & Harold Demsetz, Production, Information Costs, and Economic Organization, 62 AM. ECON. REV. 777 (1972) .................... 280
  § 5.4. Oliver E. Williamson, The Vertical Integration of Production: Market Failure Considerations, AM. ECON. REV., May 1971 (special issue), at 112 .......................................................... 292
  § 5.5. Oliver Hart, An Economist’s Perspective on the Theory of the Firm, 89 COLUM. L. REV. 1757 (1989) ................................................................. 304

§ 6. Economics and the Public-Private Choice ........................................... 316
  § 6.1. Public Ownership as Political Management ........................................... 316

§ 6.2. Public Ownership and the Theory of the Firm .................................................. 323
 § 6.2.1. Andrei Shleifer, State Versus Private Ownership, J. Econ. Perspectives, Fall 1998, at 133 .......................................................... 323

§ 6.3. Public Law and Accountability ........................................................................ 360

§ 6.4. Case Studies ..................................................................................................... 418

§ 7. Economic Theory Applied: Liability and Other Doctrines ................. 448
 § 7.1. Section 1983 and Qualified Immunity ......................................................... 448
 § 7.1.4. McKnight v. Rees, 88 F.3d 417 (6th Cir. 1996) ................................ 460
 § 7.1.5. Richardson v. McKnight, 521 U.S. 399 (1997) ................................ 466

§ 7.2. Government Contractor Immunity Under the Federal Tort Claims Act .... 483

§ 7.3. The Dormant Commerce Clause ................................................................. 494
 § 7.3.2. United Haulers Ass’n, Inc. v. Oneida-Herkimer Solid Waste Mgmt. Auth., 127 S. Ct. 1786 (2007) .............................. 506

§ 8. Vouchers and Consumer Choice .................................................................. 514
 § 8.2. School Vouchers ......................................................................................... 528


§ 8.3. The Politics of Vouchers .................................................................................................................. 565


§ 8.4. Vouchers in Other Fields ............................................................................................................. 578


§ 9. The Privatization of Force ................................................................................................................. 611

§ 9.1. General ............................................................................................................................................ 611


§ 9.2. Private Prisons .................................................................................................................................. 675


§ 9.2.4. John J. Dilulio, Jr., *What’s Wrong with Private Prisons*, PUB. INTEREST, Summer 1988, at 66 .................................................. 765


§ 9.3. Foreign Affairs Privatization ........................................................................................................ 790


Draft — do not distribute!
§ 0. Introduction


Privatization is a word with many different meanings. On a social and cultural level, it implies an individual’s withdrawal from civic life and reorientation towards the pursuit of self-interest. In the context of government, the focus here, the term is conventionally understood to signify a transfer of public responsibilities to private hands. Yet history demonstrates that increased privatization often goes hand in hand with expansion rather than contraction in public responsibilities. The government turns to private entities to provide the expertise and personnel it needs to fulfill its new tasks. Moreover, privatization is not simply a neutral phenomenon; it carries inherent political and ideological implications. Efforts to privatize public programs frequently reflect a wish to limit the government’s role and are often tied to particular substantive visions of government policies. In addition, identifying an area of activity as purely private implies that it is not an appropriate subject for public regulation or collective responsibility.

Further complicating the definitional challenge is the broad array of public-private relationships that the term “privatization” encompasses. As John Donahue has noted, “[t]wo concepts share the same word—privatization. The first concept . . . involves removing certain responsibilities, activities, or assets from the collective realm. . . . [T]he second . . . [involves] retaining collective financing but delegating delivery to the private sector.” Examples of the first type of privatization are the government selling off state-owned businesses or disbanding a government program altogether. In the United States, however, privatization overwhelmingly takes the second form. Examples here include instances when government contracts with private entities to provide goods and services for itself or others, or provides vouchers or other subsidies that allow individuals to purchase such private goods or services on their own. In some privatization contexts, the government does not provide direct funding but nonetheless uses private entities to achieve its programmatic goals—for instance, by chartering private corporations to provide desired services or relying on private actors for the content and enforcement of government regulations.

Private actors have long played central roles in government programs. But, as other commentators have noted, contemporary government privatization is in important ways a new phenomenon. Recent years have witnessed an increase in the roles private actors play in government. Both the character and quantity of private involvement are changing. Governments are giving private actors greater discretion over the implementation of government programs than in the past and utilizing new types of private partners, particularly for-profit companies and religious organizations. Privatization seems likely only to expand further in the near future, fueled by increasing belief in market-based solutions to public problems.
While private involvement in government has increased in many areas of government activity, it is particularly pronounced in social welfare programs and in government-run institutions. Four areas marked by notable expansions of privatization are Medicare and Medicaid, welfare programs, public education, and prisons. An examination of these areas is particularly instructive in demonstrating the trend towards greater privatization and how it impacts government programs.

1. Medicare and Medicaid Managed Care

One example of the trend towards expanded privatization is Medicare and Medicaid managed care. Until recently, the basic model for both programs was fee-for-service, under which the government reimburses a doctor or other medical provider for each service provided to a beneficiary. By contrast, under managed care the beneficiary enrolls with a managed care organization (MCO), usually a for-profit enterprise, and the government pays the MCO a set amount over a given period (the “capitated rate”) regardless of the medical services actually provided. While enrollment in a MCO is optional for Medicare beneficiaries, it is mandatory for many participants in Medicaid. The number of beneficiaries enrolled in managed care in the two programs has increased dramatically since 1990, with nearly 60% of Medicaid beneficiaries and 12% of Medicare beneficiaries being treated through MCOs in 2002.

Even under fee-for-service, both Medicare and Medicaid rely extensively on the private sector. Private doctors, hospitals, nursing homes, and the like provide medical care; private intermediaries process the providers’ claims for reimbursement; and review of the appropriateness of treatment decisions and quality of care is undertaken by private medical professionals sitting on peer review organizations or utilization review committees, as well as by accreditation organizations such as the Joint Commission on Accreditation of Healthcare Organizations (JCAHO). What distinguishes managed care from these other instances of privatization is that the MCO exercises a monopoly over a beneficiary’s access to health care. The MCO controls beneficiaries’ access to health services, rather than their access to payment for already-provided health services, as in the case of the private intermediaries who process reimbursement claims. Moreover, beneficiaries enrolled in MCOs generally cannot obtain government payment for services they obtain outside of the MCO’s network, whereas under fee-for-service one provider’s refusal to provide services leaves a beneficiary free to obtain subsidized services from another. Beneficiaries are limited in their ability to switch out of an MCO, and the MCO has the power to deny a desired service or procedure on the ground that it is not medically necessary or not covered under Medicaid or Medicare.

The growth of managed care in Medicaid and Medicare, as well as in the context of private employer-provided health plans, reflects efforts to cut costs. The capitated rate approach gives MCOs an incentive to control expenses and police against unnecessary treatments. Managed care can also result in improvements in health care quality: MCOs have a financial interest in providing preventative services; they are also better positioned to ensure service coordination and enforce standards for care. In the Medicare context, managed care allows beneficiaries the power to choose additional coverage or lower out-of-pocket expenses in exchange for restrictions on their choice of provider. But there are
also obvious hazards attached to the use of managed care, most significantly that MCOs also have strong financial incentives to deny coverage for medically needed but expensive treatments. Concerns over such MCO abuse have led many states to enact legislation requiring independent review of MCO denials of treatment among other measures, while the federal government has a detailed procedure providing for similar review of service denials by Medicare MCOs.

2. Welfare Privatization

Even more dramatic expansion in privatization is evident in the welfare context, which is also characterized by extensive and longstanding private involvement. Private organizations run homeless shelters and food banks; provide treatment services; operate Head Start programs; and work closely with child welfare agencies. The 1996 welfare legislation, which replaced Aid to Families with Dependent Children (AFDC) with the Temporary Aid to Needy Families (TANF) program, paved the way for greater privatization by expressly permitting private administration of state TANF programs. Moreover, TANF’s mandate that certain percentages of a state’s caseload must be engaged in a work activity means that many more welfare recipients are participating in privately run work programs or receiving services from private entities. According to a recent report, state and local governments spent more than $1.5 billion in 2001 on contracts with private entities for TANF-related services. The 1996 legislation also altered the nature of private involvement in welfare by authorizing states to provide welfare services through pervasively sectarian organizations and prohibiting governments from discriminating against religious providers, popularly referred to as “charitable choice.”

Wisconsin provides the prime example of a government putting operation of its welfare system in private hands. Private contractors administer Wisconsin’s welfare program—known as Wisconsin Works or W-2—in Milwaukee, which has approximately 75% of the state’s welfare recipients. In this role, private contractors remain responsible for substantial aspects of program administration, such as determining applicants’ eligibility for benefits, assessing their ability to work, developing employment plans, and sanctioning beneficiaries for noncompliance with program requirements. Only a few counties have followed Wisconsin’s initial lead and engaged in such wholesale privatization of welfare programs. Increasingly, however, other governments are similarly contracting out case management functions, as well as particular aspects of program administration, such as benefits delivery and child support collection. Workforce development, job training, and job placement services are especially likely to be privatized.

As in the Medicare and Medicaid managed care context, privatization of welfare administration means that private contractors wield broad authority over welfare program participants. This is true even where basic eligibility and sanctioning decisions continue to be made by government officials; for example, a private contractor’s policies as to whether beneficiaries are required to take any job offered, regardless of level of pay or work hours, can determine whether beneficiaries are referred to government agencies for sanctioning. Two developments further enhance the power of private welfare contractors.
One is a trend towards consolidating responsibility for managing programs and selecting providers in a few large contractors. The other is increasing use of “performance management,” under which the government sets performance goals and allows private entities broad discretion regarding how to achieve them. These two developments have led, in turn, to increased reliance on for-profit contractors, as nonprofits often lack the capacity and financial resources to undertake such large-scale contracts, where payment is often significantly delayed and contingent on program outcomes.

Not surprisingly, the same concern that private contractors will use their broad powers to advance their own interests at the expense of Medicare/Medicaid beneficiaries and the public also surfaces in the welfare context. For instance, under a performance-based system providing financial rewards for the number of successful job placements, private contractors have a visible incentive to try to serve only the most employable beneficiaries, or to dissuade hard-to-employ individuals from continuing in programs by means of onerous participation requirements and sanctions. But focusing solely on the potential for abuse from welfare privatization is again too one-sided. Government-run welfare programs are often characterized by abusive procedures designed simply to keep individuals off the rolls, whereas privatization may mean services are provided by nonprofits with ideological commitments more allied with beneficiaries’ interests. In addition, the operational flexibility of private providers can make them better able to improve staff performance and tailor their programs to meet the needs of particular participants or employers.

3. Privatization of Public Education

Public education is a third area characterized by recent moves to greater privatization, with an accompanying shift of core educational responsibilities to private hands. Charter schools, private management of public schools by educational management organizations (EMOs), and voucher programs provide the main examples.

Charter schools—the most significant of these initiatives to date—are publicly-funded schools allowed to operate free from many of the rules governing traditional public schools. They are also least clearly an instance of privatization; in addition to being publicly funded, they are officially denominated public schools, come into existence as a result of government authorization (the grant of a charter), and are subject to the open admissions requirement applicable to traditional public schools. Yet charter schools also embody substantial private involvement: private individuals or groups initiate the creation of the school; the schools are headed by private boards; and a significant number are managed by EMOs, usually for-profit entities. EMOs also occasionally have won contracts to operate traditional public schools, in some cases managing all or many of a district’s schools. Under voucher plans, the government provides a set amount of public funding per student to help cover tuition at private or out-of-district public schools. Overwhelmingly, students obtaining vouchers enroll in sectarian schools. Until recently, only a few publicly-funded voucher plans had been implemented. But voucher use seems likely to increase in light of the Supreme Court’s recent decision in Zelman v. Simmons-Harris (2002), which upheld Cleveland’s voucher plan against an Establishment Clause challenge.
In all three instances, private entities wield broad control over state-funded education. Charter school boards and EMOs operating charter schools possess considerable latitude over curriculum, discipline policies, and most aspects of school operation—indeed, providing this autonomy is the underlying rationale of the charter school movement. EMOs often exercise similarly broad powers when they manage public schools. Schools participating in voucher programs often are subject to only minimal qualifying requirements, and few regular mechanisms for public oversight exist. Further enhancing the power of these private entities are the significant practical obstacles that limit students’ ability to transfer schools, particularly during the school year. Moreover, here too exist concerns that the schools’ interests may not align with those of students; for example, schools receive a set amount per student, thus creating incentives for them to avoid or expel students who require more expensive educational services. An additional danger, particularly with regard to charter schools and voucher programs, is that the schools’ general freedom from oversight may lead to public funds being used to foster educational agendas that the public has refused to support.

Complicating the picture, however, is the factor of choice. Enrollment in charter and voucher schools is voluntary, and students have the option of remaining in their regular neighborhood or district school or perhaps attending public school in another district. Moreover, students usually are given the ability to transfer to another public school if their school becomes privately managed, as are teachers. The main effect of these privatization initiatives, from another perspective, is to empower parents and students, particularly given the prevalence of charter schools and voucher programs in urban areas with perennially failing public schools. Interestingly, however, parental choice also represents yet a further way in which these measures privatize public education; decisions about educational content and quality become a personal rather than collective responsibility, thereby creating schools that, in essence, are private communities of like-minded families.


1. Policing

. . . In the post-War era, public policing had significantly supplanted private policing as the main means of law enforcement and public protection in the United States. . . . [But s]ince the early 1970s, the private security sector has boomed . . . . Private security guards currently outnumber public police officers by a ratio of 3:1 in the United States. Americans now spend more than twice as much money on private security guards ($90 billion) as we spend on public police ($40 billion) every year. . . .

After the Aviation Security Act of 2001, private police no longer protect America’s airports—at least, not by themselves. But few people realize that private police protect
Almost every other kind of space in our country . . . . Almost as much as public police, private police protect public properties: In 1995, forty percent of the security work performed for federal, state, and local governments was contracted out to private firms. . . . In countless locations across the United States, private police conduct patrols, investigations, interrogations, searches, and make arrests.

. . . In the last thirty years, the institutional divisions between public and private policing have become blurry again, and they promise to become blurrier. Private police are often “deputized,” or given general public police authority, by . . . governments. Public and private police often work together on projects . . . . Public police have increasingly charged for services, demanding fees for false alarms, or offering supplemental, for-profit protection to guard local merchants, residents, and public events. Public police moonlight more than they used to, often in uniform, sometimes through the coordination of police department bureaucracies. At all institutional levels, public police are retiring to become private police, and they are retiring earlier to do so. Experience in the private security sector is regarded as a valuable asset in public police hiring and vice-versa. In a final turn, public police forces have recently begun drawing upon private contributions, and reinventing themselves on the model of private security firms.

Admittedly, there are still some important differences between what public and private police actually do. Speaking very broadly, public police are still more likely to use force, make arrests, detain suspects, and conduct searches than private police. But surprisingly few of these distinctions are formally prescribed by statutes . . . .

2. Punishment

. . . [Beginning in the 1970s,] federal and state legislatures introduced a large array of anti-crime measures that dramatically increased the frequency and length of American prison sentences . . . . The American prison population skyrocketed . . . . Today, Americans spend roughly $35 billion per year on imprisonment, and the overcrowding of our prisons grows steadily worse.

From the very beginning, the American private sector was well-positioned to take advantage of this penal revival. In the early 1970s, the federal government had already begun to support research and technical assistance that would foster public-private partnerships in prison industries, and state governments had already begun to try out such ventures . . . .

By the end of the 1980s, the private prison industry had emerged and had begun to take on the task of providing punishment in the United States. Complete private ownership and management became common in low-security facilities housing the mentally ill, illegal aliens, juveniles, and minor offenders in most states. In the 1990s . . . private corporations expanded into the business of managing minimum, medium, and maximum security adult prisons . . . . By 2000, [private] prisons housed roughly 120,000, or six percent, of the nation’s 2 million prisoners. Today, the private prison population is growing four times faster than the overall prison population . . . .
There are two American firms in the industry—Corrections Corporation of America and Wackenhut Corrections Corporation—that collectively generate about $350 million in revenues, and control about seventy-five percent of private prisons, both in the United States and worldwide. Although CCA and Wackenhut began by operating juvenile and INS facilities, the bulk of their business is now minimum- and medium-security adult prisons and jails.

In the United States, federal and state governments often enter into arrangements with private prison companies, under which the corporations raise capital, design and construct facilities, manage day-to-day operations, and sign long-term lease-purchase agreements. Recently, private companies have begun to build prisons prospectively, without prior state approval. Management contracts are usually awarded to the lowest bidder, and pay schedules are based on per inmate, per day fees. Private prison statutes typically (1) authorize public officials to enter prison management contracts with corporations; (2) describe bidding procedures, which often include mandatory savings objectives; (3) demand that the contractor carry insurance and obey all applicable laws and regulations; (4) provide for state monitoring; and (5) reserve the state’s authority to decide when prisoners are released.

3. Military Force

In comparison to the development of private policing and private prisons, the emergence of private military corporations is a more recent, less well-known phenomenon. To understand the revival of the private military market, one must first understand the business of war.

For the first half of the twentieth century, the size and strength of the . . . Armed Forces reflected America’s engagement in warfare: They increased during times of crisis and decreased during times of peace. . . . Once the Cold War settled in, however, . . . [t]he United States . . . built up unprecedented levels of peacetime readiness. . . . The demand for military equipment and supplies swelled and begat a large, permanent, private defense industry. From its very beginnings, this industry was closely allied to the Department of Defense, one of the most powerful interest groups in the United States.

I am not about to say that these early defense contractors were private military institutions. In terms of this Article, they were not military institutions at all. They did not fight wars; they produced military equipment and supplies. Military force was still “public,” still monopolized by the [U.S.] government. In fact, we might say that . . . the United States and NATO established a monopoly of military force throughout much of the world. Although some liberal citizens continued to fight as mercenaries, they fought only in small, informal groups, often in violation of domestic laws, and always without any explicit, public acknowledgement, permission, or encouragement from liberal states.

* [Wackenhut Corrections Corp. changed its name to The GEO Group in November 2003 under the terms of a share purchase agreement with another company. —Ed.]
But in 1990, things changed again. The Cold War ended. First, NATO began to
demobilize its armed forces; second, NATO became more reluctant to commit large
numbers of troops abroad for prolonged stays, especially in contested territories.

Of course, the end of the Cold War was not the dawn of world peace. Since 1990, there
have been many large, prolonged military battles in Africa, Eastern Europe, and Asia in
which NATO refused to intervene. But the United Nations, NATO, and less developed
states and resistance groups have continued to demand weapons, technology, intelligence,
strategy, training, and soldiers. Before long, “private military companies” (“PMCs” or
“private armies”) emerged to meet the military demands of less developed states and
resistance groups, filling the supply gap left by the demobilization of NATO forces.

Today’s PMCs typically offer clients a variety of military services: the training of
ground, air, and sea troops, logistical support and specialization, high-technology
intelligence systems, materiel procurement, static-site defense, and peacekeeping.
Military Professional Resources, Inc. (“MPRI”) is one of the more successful PMCs in
the United States. Based in Alexandria, Virginia, MPRI trains and equips soldiers, and
provides military expertise, planning, organization, and logistical support. Even more
than most defense contractors, MPRI is intimately associated with the United States
Department of Defense: The CEO of MPRI is a former Army Chief of Staff, and the
Board of Directors consists mainly of retired United States generals and admirals. MPRI
has hundreds of former United States special forces personnel training and equipping
foreign soldiers throughout the world. The [U.S.] government has supplied over $100
million in surplus equipment to assist MPRI, . . . refer[s] governments and organizations
seeking military training and equipment to MPRI, and reviews every MPRI contract
before it is signed. The Pentagon has contracted with MPRI to teach Army ROTC at 215
universities in the United States and to train, coordinate, and support dozens of missions
in American interests. The company offers opportunities for [U.S.-]trained soldiers to
fight in such international hotspots as Macedonia, Bosnia, Serbia, Croatia, and Africa.
MPRI has long-term contracts to supervise all aspects of several multinational United
Nations-sponsored peacekeeping teams, such as the African Crisis Response Initiative.

As you have probably noticed, I did not say that MPRI fights battles or wages wars. Like
most PMCs, MPRI claims that it does not. Generally, PMCs do everything but fight:
They train foreign troops before wars, provide back line support during wars, and
perform peacekeeping duties after wars. They tend to be rather taciturn and mysterious
when asked how often they exercise force. They are especially reluctant to say that they
initiate combat for clients. There is no question that they often find themselves in combat
zones, and on front lines, but they claim to restrict themselves to logistical and defensive
tasks. They studiously avoid labels like “mercenaries” or “soldiers-for-hire,” preferring to
market themselves as “private security firms,” staffed by officers who are
“professionals,” “leaders,” and “strategists” and soldiers who are “advisors,”
“specialists,” and “peacekeepers.” Both the industry and the clientele are committed to
guarding the secrets of particular missions. Thus, it is hard to know when, how, and how
much PMCs exercise force. But it has become increasingly obvious that some do.
Three examples are especially clear. Recently, two companies—Executive Outcomes and Sandline International—have openly engaged in direct military conflict for clients in exchange for fees. Both companies have won battles decisively in Africa. In March 1995, the government of Sierra Leone hired Executive Outcomes—a PMC based in South Africa—to win back the rebel-beleagured Kono diamond mines in exchange for the right to operate the Koidu diamond field, once the government was reinstated. Executive Outcomes took just eleven days to accomplish what the government’s armies could not: They drove the rebels away from the capital and, with the support of the company’s combat helicopters and MIG-23 fighters, chased the rebels out of the diamond fields. By the end of 1996, Executive Outcomes had beaten back the rebels for eighteen months, but the Sierra Leone Army had grown upset at the government’s continued and increasing reliance on outside armies. The government signed a peace treaty with the rebels, and Executive Outcomes was asked to leave the country in January 1997.

Just four months later, the officers of the Sierra Leone Army seized control of the country and began murdering political opponents. . . . In March 1998, the British High Commissioner in Sierra Leone, acting under the apparent authority of the British Home Office, asked a British corporation, Sandline International, to help train and equip a local force, which would be capable of removing the generals. Sandline employees trained and directed the West African multinational peacekeeping force (staffed mostly by Nigerians) and equipped them with thirty tons of small arms (imported by Sandline). Under Sandline’s direction, the West African force restored the elected president of Sierra Leone to power. When the news broke in Britain, a national political scandal erupted. The Prime Minister was nearly rebuked by Parliament, and Parliament developed a bill to regulate—not prohibit, but regulate—future dealings between PMCs and British officials.

But Colombia is the most dramatic and recent example on record. Numerous PMCs are working under contracts with the Colombian government, the [U.S.] Department of Defense, and the [U.S.] Department of State at a . . . cost of between $770 million and $1.3 billion. They make up roughly twenty percent of the American military personnel working in Columbia, flying BlackHawk attack helicopters to assist Colombian security forces in the ongoing military campaign against drug cartels and Marxist guerrilla rebels.

We should be careful not to make too much or too little of these examples. . . . They are relatively small conflicts that involve modest commitments of equipment and troops. At most, they suggest a limited trend toward a public/private substitution: In some conflicts, when modern liberal states refuse to intervene, PMCs fill the gaps. On the other hand, . . . liberal states might be responsible for these arrangements to some degree. The British government has permitted and encouraged PMCs to fight in Sierra Leone, and the United States government remains deeply involved in the day-to-day business of MPRI.

But what about the big wars, in which modern liberal states directly intervene? Surprisingly, PMCs have become increasingly involved in these conflicts too. . . . [A]n increasing number of civilian defense contractors have been joining government military personnel on the front lines. Private defense employees made up roughly one percent of personnel deployed in the first Gulf War; they were approximately ten percent of personnel deployed in Bosnia and the second Gulf War.
There have been abuses and scandals. In Bosnia, for example, seven DynCorp employees were fired for operating a sex ring, in which underage Bosnian women were held as slaves. But in the private military industry, the future looks bright. DynCorp already protects State Department Employees around the world and recently won contracts to protect the new President of Afghanistan, Hamid Karzai, and to “re-establish police, justice, and prison functions in post-conflict Iraq.” MPRI already conducts troop training throughout Africa and recently won a contract to help build a coast guard in the oil-rich waters of Equatorial New Guinea. And, while the news from Iraq is still new, one observer predicted that PMCs would play a bigger role in Iraq than they have played in any previous war. By all accounts, the revival of PMCs has barely begun.


2. The Advent of Combat-Related Privatization

Combat-related military privatization arose in the 1990s at a time when considerable cutbacks in the size of the U.S. Armed Forces were underway, when technological and geostrategic changes transformed national security practices, and when traditional types of covert operations, utilized in Southeast Asia in the 1970s and Latin America in the 1980s, had fallen into serious disfavor.

Today’s contractors, for their part, have taken considerable steps to upgrade the image of what has historically been an unsavory profession, thus helping to make the outsourcing of combat responsibilities more palatable. Indeed, contemporary American outfits are not dyed-in-the-wool bands of ruthless warriors, but rather they are incorporated businesses often headed by retired generals and colonels who have traded in their fatigues for pinstripes and left the barracks for the Beltway. Their employees, in turn, are not a ragtag lot pulled from the ranks of society’s denizens like the French Foreign Legion of yesteryear, but are likewise often recruited from among the most decorated echelons of the American military establishment.

For example, one notable contractor, MPRI, a major participant in the Balkans during the war-ridden 1990s as well as in the Latin American drug wars, boasts of having “more generals per square foot than the Pentagon.” Indeed, MPRI’s veritable “dream team” includes General Carl Vuono, former Army chief of staff during the invasion of Panama and the first Gulf War, Lt. General Harry Soyster, a onetime director of the Defense Intelligence Agency, and General Crosbie E. Saint, the former commander of the U.S. Army in Europe. MPRI advertises a breadth of competency that includes airborne operations, the provision of air support for ground troops and convoys, counterinsurgency work, force integration, tactical and strategic intelligence, reconnaissance, security assistance, and weapons control. Another contractor, SAIC, a corporate giant with annual

Draft — do not distribute!
revenues topping $1 billion, boasts of a blue-ribbon directorate that includes two former defense secretaries (William Perry and Melvin Laird) and two former CIA directors (John Deutch and Robert Gates). Other notable—and influential—firms such as Blackwater USA, DynCorp, Ronco, CACI, and Titan, are also led by former high-ranking military officers. The presence of distinguished leaders and reputable ex-soldiers impresses upon government decisionmakers that these businesses will be responsible, professional partners.

In addition to their all-star rosters, these firms have gained credibility and legitimacy because of their corporate ties. Many of the major contracting firms have close connections not only to the Pentagon but also to Wall Street, and are actually divisions or subsidiaries of such prominent businesses as Northrop-Grumman, Booz Allen Hamilton, the Carlyle Group, and Bechtel. Hence, corporate oversight and shareholder pressure may provide external sources of discipline and conformity.

3. A Survey of Recent Combat-Related Private Contracts

As mentioned above, in recent years, private military firms have protected the Karzai administration in still-unstable Afghanistan, secured American civil and military installations and served as interrogators in Iraq, bolstered and then counterbalanced the military capabilities of both the Bosnians and Croats in the Balkans, engaged in surveillance, reconnaissance, and coca-crop destroying as well as in counter-insurgency missions in Latin America, staffed security details for American officials in, among other areas, the Middle East, and attempted to bring some stability to war-ravaged Rwanda. This policy of federal contracting with private forces to serve in an array of critical zones of conflict to support American national security and foreign policy interests involves the delegation of not simply commercial responsibilities and accordingly represents a startling departure from previous partnerships with the private sector. In an effort to provide more specific details, I discuss six case studies.

a. Latin America

The United States’s lukewarm commitment to fighting the War on Drugs at its sources has set the stage for the introduction of military contractors. With stringent limitations imposed by Congress regarding the number of U.S. Armed Forces personnel and the scope of their activities in Colombia, and therefore only a relatively modest contingent of U.S. troops and officials present on the ground, the Clinton administration turned to contractors, awarding them over $1.2 billion in contract work to slow down the production and exportation of narcotics. In this capacity, private agents, notably from DynCorp and MPRI, have helped train local enforcement agents in counter narcotics work; but they have been more than just advisors: these contractors have flown sensitive reconnaissance missions, patrolled the skies to turn back (under the threat of force) smugglers, and piloted crop-dusters to destroy coca fields. Their efforts have not gone unchallenged and, as a result, military contractors have at times been drawn into firefights with narco-traffickers and even leftist rebels, some of whom had no direct connection to the drug trade.
In the course of their dangerous work, a number of American contractors have been killed; these casualties have largely escaped public notice, media attention, and congressional scrutiny. Indeed, relatively little is known about the extent of America’s involvement in Colombia, let alone details regarding the delegation of specific activities to private firms. And, although the GAO rated DynCorp’s performance in Latin America as “unsatisfactory” over several years, the State Department repeatedly renewed the firm’s contract.

b. The Balkans

In the Balkans during the mid-1990s, the bloody contests between and among Serbs, Croats, and Bosnian Muslims produced unspeakable carnage and threatened to destabilize the entire region. The Clinton administration, hamstrung by U.N. arms embargos, hesitant allies, wary adversaries—not to mention internal White House indecision and congressional opposition and, also, its desire to retain the appearance of an honest, neutral broker in the region—was militarily limited in its ability to help quell the violence. Nevertheless, the Administration actively wanted to resolve the conflicts and chose, in part, to augment the relative military strength and self-sufficiency of the Croats and, later, the Bosnian Muslims to counter Serb aggression.

Unable, for the reasons mentioned above, to provide direct assistance through much of the years of fighting (but also unwilling to remain fully on the sidelines), the United States turned to private solutions. First, it sought to bolster the fledgling Croatian state and arranged for the American firm, MPRI, to provide strategic and tactical military training as well as instruction in modern weaponry. In working “under the guise of a private commercial enterprise, MPRI could achieve what would otherwise be impermissible military objectives.” Since directly supplying training and materiel to the Croats would have violated the U.N. arms embargo, and perhaps prompted Russia, in turn, to fortify its traditional ally, the Serbs, the United States’s use of MPRI effectively permitted it to remain neutral yet still pursue its unilateral humanitarian and geostrategic interests in the region.

Then, later, to entice the Bosnian Muslims to accept the Dayton Peace Accords, the need arose to strengthen their position, too, vis-à-vis the Serbs. Again, the United States—intent on remaining ostensibly neutral—played matchmaker and, interestingly, recommended MPRI’s services. As a matter of fact, the Bosnians ultimately conditioned their signing of the Dayton Accords on the State Department’s promise to secure for them “the same guys who helped the Croatians.” So, while the United States committed thousands of troops to the region as neutral peacekeepers (to enforce the Dayton agreement), it also helped the Bosnian Muslims acquire additional support: Privatized intervention thus allowed Washington to have its cake and eat it too.

In both Croatia and Bosnia, the training allegedly exceeded what one might expect a purely advisory engagement to entail. In fact, some reports of the contractors’ involvement invited comparisons to what had transpired in the early years of America’s “advisory” involvement in Vietnam. The training in the Balkans included practical instruction such as strategic planning and targeting enemy locations, skills that were soon
utilized in actual offensives. In one particularly bloody campaign, in which the Croatian leaders in command were ultimately charged with international war crimes for their brutality, it has been alleged that MPRI was intimately involved in all stages of planning.

Fortunately, while the participation of such advisors did not lead to an escalation of America’s entanglement, as was the case in Vietnam, the story of private soldiers in the Balkans nevertheless gets worse. DynCorp, the same company employed to protect President Karzai, the same company that received unfavorable performance ratings in Latin America, and the same company that is now spearheading a good deal of security-oriented contracting work in Iraq, was (along with MPRI) also providing security services in Bosnia. While there, DynCorp personnel were accused, by colleagues and by the British government, of operating a full-fledged sex-slave operation involving young female war refugees. Given the vagaries of the contractors’ legal status and the jurisdictional limitations of American criminal law, there was little the United States could do, that is, short of refusing to contract for DynCorp’s services in the future. As explained above and below, however, the United States has not even taken that modest step.

c. Afghanistan

As referenced in the Introduction, in the Fall of 2002, the United States withdrew its elite Special Forces team assigned to protect President Karzai and, in its stead, contracted (yet again) with DynCorp to provide security. Ensuring the stability and safety of the pro-Western Karzai regime, I need not add, is widely considered absolutely critical not only to rebuilding a free Afghanistan, but also to waging a successful war on global terrorism. Nevertheless, though the decision to privatize came at a time when Kabul remained incredibly unstable and threats on the new president and his cabinet were tangible and ever-present, Defense Secretary Rumsfeld insisted that privatization was a necessity: He simply could not spare the handful of troops any longer.

This justification may not seem totally satisfying. The military detail originally assigned to Karzai numbered approximately forty Special Forces soldiers. To put that number in perspective, conservative estimates suggest that, at the time, the total number of active U.S. Special Forces personnel was between 40,000 and 50,000 strong. And, moreover, there were tens of thousands of additional regular American soldiers stationed throughout Afghanistan carrying out all sorts of duties, from protecting the construction workers building roads to rooting out Taliban and al-Qaeda operatives in the caves along the Pakistani border. Finally, as mentioned above, DynCorp has received abysmal performance evaluations ranging from poor service in Latin America to horrible human rights violations in Bosnia. More recent reports, from Fall 2004, have described DynCorp employees as alienating and intimidating locals in Kabul. Nevertheless, despite the obvious significance and importance of protecting Karzai and the apparent option of diverting a handful of regular U.S. soldiers to relieve the outgoing Special Forces team, the Bush administration preferred this private alternative.

As an additional note regarding contractors in Afghanistan, it has also recently come to light that private contractors working as interrogators in American military prisons in
Afghanistan have been deemed responsible for brutal beatings (and even deaths) of al-Qaeda and Taliban inmates. There is even evidence of Americans running “private” detention centers, possibly—but not definitively—in some loose affiliation with the CIA, purportedly to acquire information regarding terrorism.

d. Iraq

With hundreds of thousands of contractors involved in the liberation and occupation in Iraq, in jobs ranging from cooking and construction to armed security and intelligence, no combat venue has witnessed a greater influx of American private agents. Among them, many perform traditional commercial services. But a sizeable number, estimated between 20,000-30,000 contractors, carry out many of the core security functions typically understood to be inherently governmental—and inherently soldierly. The difficulties of the occupation, coupled with the relative shortages of U.S. troops, an unwillingness to contemplate a military draft, and only minimal assistance from foreign allies have made contractors close to indispensable. Along the way, of course, many contractors have been killed. Casualties among contractors, to date, are not insubstantial, but of course they are not as high as the number of reported casualties among members of the U.S. military. Yet comparatively speaking, rarely are those contractor-casualty numbers tallied with such care, publicity, and despondency as soldier-casualties are.

In the interests of providing some descriptions of the type of private military-security work undertaken in Iraq, I offer three representative illustrations.

First, both the Coalition Provisional Authority (“CPA”) and the U.S. government have contracted with private military firms to provide security for key American and CPA positions, important Iraqi locations (such as banks, museums, and oil fields) as well as for American and CPA officials, including Ambassadors Paul Bremer and John Negroponte. These contractors often carry automatic weaponry and, at times, have been provoked into exchanging fire with insurgents. For example, in early April 2004, Shiite militia forces attacked the CPA’s headquarters in Najaf. Eight employees of Blackwater, unaided by members of the U.S. military—or by any other national army participating in the liberation and occupation—had to fend off the siege until they were ultimately supported by reinforcements. The cavalry, so to speak, came by way of a helicopter crew, comprised of additional Blackwater agents, not American military personnel. Similar battles, waged principally by armed contractors (with more or less success), have been fought in such places as Mosul, Kut, and Fallujah since the occupation began.

Second, private contractors have assumed now infamous roles as intelligence agents, translators, and supervisors in Iraq’s most notorious prison, Abu Ghraib. In the weeks following formal investigations by General Taguba, General Fay, and former Defense Secretary Schlesinger, it became apparent that employees of Titan and CACI, who devised interrogation techniques and supervised the military police, were central participants in the horrific prisoner-abuse scandal.

Third, contractors working for DynCorp helped stage and raid Ahmed Chalabi’s personal compound as well as his offices at the Iraqi National Congress in Baghdad. This raid—
which occurred soon after the Pentagon suspected that Chalabi had passed along U.S. national security secrets to Iran—is indicative of the fact that military contractors in Iraq have undertaken offensive missions. Within the industry, which vociferously contends that it only accepts “defensive” assignments, this example signals a major evolution in contractor responsibilities and protocols.

With extreme stress on the active U.S. Armed Forces, the withdrawal of troops by Coalition partners, a lack of faith in Iraqi security teams, and no end in sight to the insurgents’ hostilities, one would have to assume that the demand for (and utility of) military contractors, in spite of the notoriety they received at Abu Ghraib, will only increase.

e. Rwanda

Another interesting but not widely reported case of military privatization involved the United States supporting the very limited use of private agents in Rwanda. In the midst of that horribly brutal genocide campaign, an extremely small (and admittedly insignificant) group of private agents under the employ of the Ronco firm were dispatched to protect some villages, to provide some humanitarian relief, and to offer training to the fledgling Rwandan Patriotic Army. Contrasting the magnitude of the travesties against the modest deployment of private agents, it is safe to conclude that Ronco did not make much of a dent in stopping intertribal violence. It is even safer to say, that the United States, like most other nations, did almost nothing else to stop the genocidal massacre.

Nevertheless, despite its extremely limited scope and even more limited success, the Rwanda-Ronco project provides a powerful if incomplete model of possibilities. Irrespective of any U.N. hesitancy, it is doubtful that the American public would have countenanced U.S. servicemen and women being sent to Central Africa to stop internecine tribal violence—especially on the heels of the recent debacle in Somalia. On the other hand, the public might be more comfortable with—or less aware of—dispatching contractors, who specifically agreed to sign up for such a dangerous mission, than with sending over regular U.S. soldiers whose defense of American sovereignty and interests does not (as the public might believe) legitimately extend to humanitarian police actions. Dangerous humanitarian work such as what may be warranted today in the Darfur region of Sudan may, accordingly, prove to be a new growth industry of assignments for contractors who consent to enter dangerous situations well outside of the scope of what is conventionally understood as core American national security interests.

f. Gaza Strip

A final, recent illustration of military privatization on a very small scale involved the terrorist attacks on U.S. consular attachés in Gaza. In October 2003, a caravan of U.S. diplomats was shepherded through a virtually lawless area of Palestinian-controlled Gaza by DynCorp security forces, not State Department Diplomatic Service agents or U.S. Marines, who otherwise often guard embassies and overseas diplomats. The killing of three American contractors on the security detail did make the news for a day or so, but it did not become a serious media or diplomatic story, and little was ultimately made of the
attack in terms of creating an impetus for a counter-strike, or even a rethinking of U.S. Middle East policy. Perhaps, for better or worse, if it were American soldiers killed, a different response would have been forthcoming.


The degree to which states and international organizations have farmed out foreign affairs activities to private actors is truly breathtaking. It would not be an exaggeration to say that the U.S. government has hired private entities to help fight our wars, to deliver much of our foreign aid, and to play an important role mediating our conflicts and engaging in other diplomatic activities. Many other states are pursuing a similar course. International organizations are likewise turning to private entities to support peacekeeping and to deliver all manner of humanitarian, development, and post-conflict reconstruction assistance, as well as to participate in peacemaking negotiations.

The privatization of military functions is perhaps the most remarkable example. Probably no function of government is deemed more quintessentially a “state” function than the military protection of the state itself, and some scholars of privatization in the domestic sphere have assumed that the military is one area where privatization does not, or should not, occur. Indeed, some have argued that, by hiring private armies to keep itself secure, the state would threaten its own existence because it would have no way to control these private military actors.

Yet, governments around the world, including the United States, are increasingly hiring private military companies to perform core military functions. For decades, of course, the U.S. government has entered into agreements with private companies to build weapons and other equipment, as well as to provide the basic goods necessary to run a government agency—everything from desks to office supplies. However, such contracting activity now covers services to active troops in the field. These activities include not only support services, such as food, accommodations, and sanitation for troops on the battlefield, but also core functions such as intelligence gathering, communications, weapons maintenance, and even troop training. According to one commentator, “while contractors have long accompanied U.S. armed forces, the wholesale outsourcing of U.S. military services since the 1990s is unprecedented.”

Indeed, if one looks to U.S. forces deployed on the battlefield, the ratio of private contractors to troops has increased dramatically in the past fifteen years. In the first Gulf War, the ratio was roughly one to one hundred; in the current war in Iraq, the ratio is one to ten. Although the United States has not yet used the employees of private companies in actual combat roles, it has deployed them to fulfill tasks such as military intelligence gathering, troop training, weapons maintenance, and support functions that are very close to combat; these private actors even have the power to wield force in a variety of
circumstances. Other countries, such as Sierra Leone and Angola, have explicitly hired private armies. In many modern conflicts, these private military companies have played a decisive role.

The pervasiveness of the U.S. military’s use of contractors captured media attention with news of their role in Iraq. When stories surfaced that U.S. military personnel had abused detainees at Abu Ghraib prison in Iraq, it soon became clear that private contractors employed by CACI, Inc. and working under an agreement with the Department of the Interior had participated in the abuse. Indeed, intelligence operatives may actually have given orders to uniformed military. Translators hired under a similar contract with the firm Titan, Inc. were also implicated in the abuse. News of gruesome security contractor killings by Iraqi insurgents has sparked additional popular attention.

Yet, CACI is not an anomaly. Firms such as Kellogg, Brown, and Root (KBR) have for years built and maintained military bases, transported troops and equipment to and on the battlefield, repaired and maintained roads and vehicles, distributed water and food to troops, washed laundry, refueled equipment, attended to hazardous materials, and performed related environmental services, earning roughly $1.7 billion annually from military work. Beyond this logistical support, other companies provide core military functions such as troop training and intelligence gathering. For example, since 1996, another U.S company, Military Professional Resources Incorporated (MPRI) has run the ROTC training programs at universities around the country and has played a key role in numerous programs to educate U.S. forces, including officer training, war gaming, and tactical planning. In recent years, moreover, MPRI has expanded its services to a wide variety of countries, including Croatia, Bosnia, Angola, Saudi Arabia, Sri Lanka, Nigeria, and Equatorial Guinea, as well as to regional intergovernmental programs such as the African Crisis Response Initiative. Finally, and most controversially, military companies have provided direct combat services. For example, the now-dissolved South African company Executive Outcomes, which drew its personnel largely from the apartheid-era South African Defense Force, won contracts with governments in Angola, Sierra Leone, Uganda, Kenya, South Africa, Indonesia, Congo, and others to engage in direct combat during the 1980s and 1990s. Indeed, its activities in Sierra Leone and Angola are widely believed to have altered the outcome of the conflicts in those states.

Although privatization in the military context has received far more attention, overseas aid is another area in which states are also increasingly turning to private contractors to fulfill functions formerly performed directly by the state. From emergency humanitarian relief, to long-term development assistance, to post-conflict reconstruction programs, private actors under contract with the United States, other governments, and international organizations are taking a larger and larger role. The most dramatic surge in privatized aid has involved humanitarian relief. The United States, for example, has contracted with private companies such as KBR to build refugee camps, and with nonprofit NGOs such as Save the Children to deliver relief supplies and medical services. For fiscal year 2003, the USAID Office of U.S. Foreign Disaster Assistance spent sixty-six percent of its nearly $300 million budget through NGOs. Other countries and international organizations are similarly turning to NGOs to deliver humanitarian aid. Longer-term development aid has followed a similar trend. By the mid-1980s, development agencies
had begun to shift their focus from general funding for foreign governments to more targeted direct support both to grassroots organizations helping to eradicate poverty and to other civil society institutions seen as necessary for democracy and development. In the United States, for example, the government now uses both international and foreign NGOs to deliver much of its aid overseas, rather than providing aid directly to foreign governments. As the cases of Iraq and Afghanistan make clear, privatization is also taking place in the arena of post-conflict reconstruction, with multimillion dollar contracts awarded not only to nonprofit organizations but to for-profit corporations as well. Indeed, so far in Iraq USAID has awarded fifteen contracts worth a total of $3.2 billion to for-profit companies, while it has awarded only six grants worth $40 million to nonprofit organizations.

In addition to military and foreign assistance functions, governments have also turned to private entities to assist in peacemaking and other diplomatic tasks. The Carter Center, probably the best-known organization in this field, has engaged in high-level diplomatic efforts to avoid or end conflicts at the behest of governments around the world. For example, in 1994, former President Carter assisted the U.S. government in reopening talks with North Korea, negotiating an agreement for the peaceful departure of Raoul Cedras from power in Haiti (thereby averting the need for increased U.S. military intervention), and also brokering a ceasefire in Bosnia. More recently, the Center has engaged in peacemaking activities at the request of numerous governments, including Uganda, Sudan, and Ecuador. Other organizations have performed similar roles around the world.

Finally, it is important to note that, in addition to states, international organizations such as the United Nations have also turned to private entities. For example, the office of the United Nations High Commissioner for Refugees (UNHCR) has entered partnerships with hundreds of NGOs around the world for services including refugee protection, community services, field security, child protection, engineering, and telecommunications in emergency relief situations. Although the United Nations has not deployed private military firms in combat roles, some policymakers and commentators have suggested that such a step would provide badly needed help to peacekeeping and peace enforcement operations. And even those who do not endorse a direct combat role for private firms nonetheless argue that the United Nations should privatize military logistics functions, much as the U.S. government has done.

To be sure, one might argue that privatization in all of these areas of foreign affairs is nothing new. With respect to the military, mercenaries (loosely defined as soldiers working for private gain) have appeared throughout history. From the Swiss guards of the Middle Ages, to the merchant-armies of the British and Dutch East India Companies during the colonial period, to the privateers of early America, to the French Foreign Legion still active today, mercenaries have played a significant role over the centuries. Indeed, before the Treaty of Westphalia, which marked the emergence of the state system that eventually gave rise to large standing national armies, mercenaries were the norm rather than the exception. By the twentieth century, however, apart from some post-colonial wars of independence (where mercenaries often fought against national liberation movements), the bulk of military security work has been performed by
professionalized, bureaucratized armies and not private actors. It is against this backdrop that we have seen, over the past two decades, the increasing re-privatization of military functions.

Likewise, the privatization of foreign aid and diplomacy are not wholly recent developments. With respect to foreign aid, private groups from the United States funneled aid overseas for specific causes long before the government developed foreign assistance programs. Indeed, the practice of government-sponsored foreign assistance did not develop in a significant way in the United States until the initiation of the Marshall Plan during the period following World War II. However, from the 1950s through the 1970s, much of the aid consisted of direct grants to needy countries. In contrast, as noted previously, over the past two decades the government has delivered more and more foreign aid through nongovernmental actors. And with respect to diplomacy, private organizations have long played a role in ongoing peace negotiations and other efforts, but the high-level diplomatic work of organizations such as the Carter Center is new and distinctive.

The forces driving this trend toward privatization are not fully understood, but the dominant rationale articulated by most scholars of the subject is the promise of cutting costs. The government need not offer pensions or benefits to employees of private companies working under contract, and it can hire contractors on a short-term basis, thereby decreasing the size of government bureaucracies. Moreover, unlike many governmental employees, private contractors are typically not unionized. The lack of unionization fuels the political support of those on the right, and, at least in the United States, there tends to be broad bi-partisan support for any trend that seems to make government “smaller” and “more efficient.” In addition, in the case of the military, private military companies may offer governments greater flexibility. Such companies are touted for their ability to work quickly, and in states with ill-equipped, poorly functioning militaries, private companies can provide badly needed expertise to help train, or even replace, government troops. Using private military companies can also offset shortages of troops by offering a rapid means of growing a state’s military capacity, and states can deploy their forces with lower domestic political costs because fewer uniformed troops are put at risk, thus keeping official casualty figures down. Finally, some have suggested that the growth of private military firms has been fueled in part by the labor pool created as many military dictatorships and their accompanying security forces have been dismantled during transitions to democracy.

Foreign aid privatization also appears to be motivated largely by a desire to cut costs. Certainly in the United States the outsourcing of aid is linked to the political push for smaller government, combined with weak political support for foreign aid generally. Indeed, USAID, perhaps motivated in part by the need to justify its activities to an increasingly skeptical Congress, was one of the first agencies in the United States to take then-Vice President Al Gore’s “Reinventing Government” message to heart, declaring in 1994 that “USAID is now fully committed to reinventing itself as a more efficient, effective, and results-oriented organization.”
The privatization of diplomatic tasks such as peacemaking has received even less scholarly attention than other forms of foreign affairs privatization, and the reasons behind this trend are thus even less clear. It appears, however, that the growing use of private entities in this arena has stemmed from the high-level experience of those such as former President Carter who have founded and worked for such organizations, as well as the organizations’ independence, which frees them from some of the political costs of sensitive diplomatic efforts.

Thus, the recent rise of privatization does represent a shift, at least as compared to the recent past, away from the large, highly-bureaucratized state. Just as states are outsourcing their domestic functions, they are also outsourcing their foreign affairs activities. And while the surge in foreign affairs privatization raises many questions that merit further study, a central issue presented by this trend is whether increased outsourcing undermines the public law values that apply primarily to state actors.


Among the least noticed of the plentiful witnesses at the murder trial of O.J. Simpson was Susan Silva, then director of administration at Westec Security, Inc. The prosecution called Silva to describe for the jury certain security devices Westec had installed in Simpson’s house. Asked to describe Westec itself, Silva explained, “We’re in the business of installing, monitoring, providing service and patrolling alarm systems.”

Westec is not always so modest. Promotional brochures bill the company as the nation’s “largest full-service security provider,” offering not only alarms but “sentry officers,” “escort services,” “roving patrols” in “highly visible” vehicles, and “rapid deployment” of “highly trained armed personnel,” prepared “for virtually any situation in the field.” Outside Simpson’s house, three Westec signs warned of “armed response.”

Drive for ten or fifteen minutes through any residential part of West Los Angeles and you likely will pass a hundred or more of these Westec signs, along with a comparable number of similar placards announcing the armed protective services of Bel-Air Patrol or Protection One, Westec’s two largest local rivals. Over the past twenty years, the home security business has grown dramatically in Southern California. A decade ago Westec signs already seemed almost “as common as weeds on Southland lawns”; since then Westec has tripled its local clientele. Just within the city limits of Los Angeles—which exclude, inter alia, the affluent communities of Beverly Hills, Manhattan Beach, and Santa Monica—about 800 private guards patrol residential neighborhoods, more than one-tenth the entire patrol strength of the Los Angeles Police Department.

Private residential patrols remain much less common outside Southern California, but they are growing in popularity throughout the country. Increasingly, moreover, private guards patrol not just the properties of individual customers but entire communities.
Some of this is the result of the surge in walled-off, gated housing developments, which commonly use homeowner fees to pay for patrols by private security guards. As many as four million Americans may already live in these enclaves, and that number is growing rapidly; in Southern California, an estimated one-third of all new communities are gated. But even ungated communities are increasingly hiring private patrols, either with homeowner fees or with government-approved special assessments.

Residential security guards, moreover, are but one part—a relatively small but rapidly growing part—of the much larger workforce of private police personnel. That larger workforce has itself grown substantially faster over the past quarter century than both the population and the ranks of public law enforcement, effecting “a quiet revolution in policing.” At this point, security guards in the United States actually outnumber law enforcement personnel; there are roughly three private guards for every two sworn officers. In California, the ratio is well over two to one.

These figures are necessarily imprecise, for at least two separate reasons. The first, to which I will return later, is empirical: reliable information on the size and composition of the private security industry is notoriously sparse. The second reason is definitional. One of the hallmarks of private security is “its non-specialized character,” its tendency to be implemented in part by employees—such as store clerks, insurance adjusters, and amusement park attendants—whose principal duties at least ostensibly lie elsewhere. Still, there can be no doubt that specialized private security personnel play a large and growing role in policing America. Uniformed private officers guard and patrol office buildings, factories, warehouses, schools, sports facilities, concert halls, train stations, airports, shipyards, shopping centers, parks, government facilities—and, increasingly, residential neighborhoods. On any given day, many Americans are already far more likely to encounter a security guard than a police officer; in the words of one industry executive, “[t]he plain truth is that today much of the protection of our people, their property and their businesses, has been turned over to private security.”

Nor is private policing limited to uniformed security guards. America has over 70,000 private investigators and over 26,000 store detectives; together these individuals outnumber FBI agents by almost ten to one. The ranks of private investigators, in particular, have swelled in recent years, growing by nearly 50% during the 1980s. Private detectives increasingly are hired not only to watch for shoplifters, but also to investigate, and not infrequently to spy on, everyone from insurance claimants and litigation opponents to employees, business partners, and even prospective neighbors.

In addition to these wholly private personnel, an estimated 150,000 police officers moonlight as private security guards, often in police uniform. This practice, too, appears to have escalated sharply; more than half of the officers in many metropolitan police departments now supplement their income with private security work. In a growing number of cases police departments themselves contract to supply their personnel to groups of merchants or residents, and then pay the officers out of the proceeds.
This public-to-private personnel leasing is the mirror image of the much larger practice of out-contracting, in which government agencies hire private security companies to perform work previously carried out by law enforcement officers. According to one estimate, the fraction of security work contracted out by federal, state, and local governments increased from 27% to 40% between 1987 and 1995. The trend seems likely to continue. Much out-contracted security work consists of parking enforcement, traffic direction, and other tasks unlikely to bring the private employees into contact with the criminal justice system. Increasingly, though, government agencies are hiring private security personnel to guard and patrol government buildings, housing projects, and public parks and facilities, and a small but growing number of local governments have begun to experiment with broader use of private police. A few municipalities have hired private security companies to provide general patrol services; more commonly, groups of residents or business owners in particular areas have received permission to tax themselves (and their dissenting neighbors) to pay for private patrols.

If the precise size and contours of our national private police force remain unclear, its activities are even foggier. Security firms regularly assure the public that private guards, unlike law enforcement officers, serve only to observe and to report. But clients often receive a different message. Protection One, for example, bills itself to customers as “your personal police force”; its competitors make similar, if more circumspect, claims. About the actual activities of private security personnel there is little reliable information. Plainly, though, they often do a good deal more than observe and report. Generalizations beyond that are difficult, partly because the industry is secretive, and partly because security personnel do not all perform the same functions. Store detectives, for example, make frequent arrests; residential security guards typically do not. Even residential security guards, however, may carry out brief detentions with some frequency. And the security industry as a whole probably carries out significantly more stops, searches, and interrogations than is often imagined. More generally, despite the frequent claim by security firms that they are not “playing police,” large numbers of private security personnel today appear chiefly engaged in what is, in essence, patrol work—work once understood as the principal function of public law enforcement. “[T]he modern . . . security guard’s job, while different from that of modern public policemen, is very much like that of the traditional ‘cop on the beat.’”

Not coincidentally, the traditional “cop on the beat” has been something of a vanishing species for much of the past half century; this much-bemoaned development plainly has had more than a little to do with the proliferation of private guards during the same period. The past few years, in fact, have seen increasing calls for a revival of traditional beat policing—calls, in other words, for police officers to act more like security guards. The response to these calls has been limited, in part, by the hard fact that police patrols are expensive. Even a greatly accelerated revival of beat policing, though, would be unlikely to reverse the massive growth of private security, because, as I discuss later, changing patterns of public policing have not been the only factor contributing to that growth.
Indeed, one of the most striking aspects of police privatization is its “pervasive, international character.” The exponential growth of private security in the United States has been mirrored in Canada, the United Kingdom, Australia, New Zealand, and, to a lesser extent, the rest of the world. Private security is global in another sense as well: ownership and operation of the industry is increasingly multinational.

Other than its international character, however, little about private policing is entirely new. As we shall see, it was not until the nineteenth century in England and America that private law enforcement, and amateur, quasi-public constabularies and night watches, gave way to organized public policing; “[t]he modern police are indeed a ‘new police’ and in some important respects it is they and not the modern phenomena of agencies within the private security sector that are out of step with the historical lineage of policing forms.” Even professional, bureaucratic policing was pioneered by private firms: Allan Pinkerton founded his detective company before the Civil War, and well into the twentieth century he and his competitors provided America’s only national law enforcement organizations.

The shared, complex history of public and private policing—traced in Part II of this Article—bears heavily on contemporary understandings of the proper roles for the public and private sectors in enforcing law and maintaining order. Those understandings, however, are murkier than sometimes thought, and that murkiness is reflected in the legal regime under which the private security industry operates.
§ 1. The Non-Delegation Doctrine


... [P]rivatization is poorly characterized as government withdrawal or disinvolvement from an area of activity. That description fails not only because privatization so often accompanies an expansion in government responsibilities, but more importantly because it misses privatization’s core dynamic. In many instances of privatization, the overall context remains one of significant government endeavor; ... the government provides the funds, sets programmatic goals and requirements, or enacts the regulatory scheme into which private decisionmaking is incorporated. But the government relies on private actors for actual implementation. Rather than government withdrawal, the result is a system of public-private collaboration, a “regime of ‘mixed administration’” in which both public and private actors share responsibilities.

Viewing privatization in this way highlights how it serves to delegate power over government programs and regulation to private actors. Indeed, “in many cases the major share ... of the discretion over the operation of public programs routinely comes to rest not with the responsible governmental agencies, but with the third-party actors that actually carry the programs out.” The recent move to expanded privatization, with grants of even greater discretionary authority to the government’s private partners, enhances the extent of delegated authority. More importantly perhaps, these delegations of discretion are unavoidable because the power to implement and apply rules is inseparable from the power to set policy. While some contend that government can privatize implementation while retaining control over governance and policy management, in practice such a divide rarely exists. “Executive action that has utterly no policymaking component is rare ... .” Close government oversight or specification of policies and procedures can limit the extent of discretionary authority delegated to private actors, but cannot eliminate it. Moreover, if anything there seems to be a move towards expanding the discretionary powers of front-line workers, both public and private, with many scholars condemning the rule-bound and centralized character of contemporary regulatory regimes and urging greater flexibility in implementation.

Through their control over government programs, private actors necessarily obtain control over program participants. ... [T]his control is enhanced when private entities have a monopoly or quasi-monopoly over access to government-subsidized services or broad powers over how government institutions operate. Another factor enhancing private power over participants is that privatization frequently occurs in contexts marked by relations of dependence, in particular social welfare and human service programs. Those implementing such programs, whether public or private, gain power over program participants by virtue of their control over vital resources, as well as their greater knowledge and expertise. Dependency in turn reinforces the discretionary powers of
service providers, as often the services provided in dependent contexts defy easy or clear specification.

. . . [T]he powers exercised by private entities as a result of privatization often represent forms of government authority, and . . . a core dynamic of privatization is the way that it can delegate government power to private hands. Identifying what constitutes government power is a notoriously hazardous enterprise, and little agreement exists on where the boundaries of government power as opposed to private power lie. Interestingly, however, the powers at issue in many instances of privatization are those we conventionally think of as governmental. Few deny that private prisons are wielding government power, given that the right to physically constrain and coerce others is ordinarily reserved for the state. When private regulators determine the content and enforcement of standards governing a field of activity, their decisions similarly represent government power in the form of nonconsensual exercises of authority over others. The claim that privatization leads to private exercises of government power also seems relatively noncontentious when private entities function much like government employees, merely applying government-generated policies and requirements over which they exercise little or no independent judgment.

On the other hand, it is less evident that government power is involved where private entities are undertaking tasks they traditionally have performed and over which the government allows them substantial discretion. Instead, private entities’ ability to shape government programs in such contexts seems just an indirect side-effect of their autonomous determinations, often made pursuant to independent professional standards. One argument as to why private actors are wielding government power even in these contexts is that their positions in government programs give them control over program participants. Yet private actors often wield similar powers when they act independently of government; for example, MCOs exercise the same control over access to health care for participants whose coverage comes privately, through employers or individually purchased health insurance, as they do for those whose coverage is publicly subsidized. Thus, the fact that private actors exercise power over vulnerable third parties, absent more, would be insufficient to distinguish the powers being exercised in these contexts as uniquely governmental.

In many instances of privatization, however, something more is involved. Private providers and regulators do not simply wield power over others, but they also control third parties’ access to government benefits and resources. By virtue of their role in government programs, therefore, private entities gain a distinct “mantle of authority” that enhances their ability to cause harm. This control over access to government benefits is sometimes clear and direct; for example, when private actors determine whether an individual is eligible for government-funded services. More commonly, however, this control is exercised indirectly, through the private entities’ policies regarding the substance of the services they provide or their decisions with respect to a particular participant’s needs. Whether private actors are applying government-specified criteria or instead making independent determinations affects the direct or indirect character of their control, but does not alter the underlying dynamic. Thus, denials of medical services by Medicare and Medicaid MCOs control beneficiaries’ access to government-subsidized
health care even where these denials rest on professional medical standards rather than
government rules regarding covered procedures. Private contractors similarly exercise
control over access to publicly funded services when they assess a welfare participant as
not needing additional job training, as do private schools in voucher programs when they
suspend participating students for disciplinary infractions.

When exercised by public actors, control over government resources is commonly
thought of as government authority, and transfer of such control to private individuals
and entities does not inherently change the nature of the power at issue. Of course, in
some instances the transition from public to private performance does impact the nature
of the power being exercised. It seems intuitively implausible that decisions by job
placement counselors at a nonprofit organization are exercises of government authority
simply because the organization’s job placement services are funded by a government
grant. Instead, whether such private counselors wield government authority turns on their
relationship to the government. Public-private relationships range across a wide
continuum . . . . Insofar as the government funds the nonprofit simply to foster provision
of job placement services in a community, then the private counselors’ activities appear
genuinely independent and nongovernmental. But if the government uses the nonprofit to
implement its welfare program—for instance, requiring that welfare beneficiaries
participate in the nonprofit’s job placement program as a condition for receipt of income
assistance—then the distinction between the private counselors and their public
counterparts in the state welfare department begins to disappear. Another way to put this
point is that in the former case, the government is not using the nonprofit to control third
parties’ access to government benefits; instead, the nonprofit itself is the direct and
intended recipient of government aid.

The expanding privatization documented above means that government increasingly is
using private actors to operate government programs in its stead, and, as a result, private
entities are wielding even greater power over participants’ access to government
resources. Simultaneously, however, the broad discretion granted to private entities also
supports seeing the underlying program as being one where the government seeks to
foster independent private action. Line-drawing is thus very difficult. One point to
emphasize, however, is that individuals’ entitlement to government benefits or to
participate in a particular program has no determinative significance on the question of
whether privatization involves a delegation of government power. For example, the
conclusion that a hospital has no entitlement to participate in Medicare does not, in and of
itself, preclude finding that the government has delegated government power if it makes
JCAHO accreditation decisions determinative of the hospital’s eligibility to participate.
Such entitlements are relevant, to be sure, but primarily in terms of what they
demonstrate about the government’s intent in engaging in privatization; the absence of an
entitlement is one factor that may suggest the government is seeking to foster purely
independent private action rather than rely on private entities to more directly serve its
own programmatic goals. But it is the use the government makes of the private entities
involved, rather than the rights or status of beneficiaries and other program participants,
that is the ultimate focus.
§ 1.2. Panama Refining Co. v. Ryan, 293 U.S. 388 (1935)

Mr. Chief Justice Hughes delivered the opinion of the Court.

. . . [T]he President, by Executive Order, prohibited “the transportation in interstate and foreign commerce of petroleum . . . produced or withdrawn from storage in excess of the amount permitted to be produced or withdrawn from storage by any State law . . . .” This action was based on § 9(c) of Title I of the National Industrial Recovery Act . . . [, which] provides:

“The President is authorized to prohibit the transportation in interstate and foreign commerce of petroleum . . . produced or withdrawn from storage in excess of the amount permitted to be produced or withdrawn from storage by any State law . . . .”

Fourth. Section 9(c) is assailed upon the ground that it is an unconstitutional delegation of legislative power. The section purports to authorize the President to pass a prohibitory law [related to] the transportation in interstate and foreign commerce of petroleum and petroleum products which are produced or withdrawn from storage in excess of the amount permitted by state authority. . . . [W]e look to the statute to see whether the Congress has declared a policy with respect to that subject; whether the Congress has set up a standard for the President’s action; whether the Congress has required any finding by the President in the exercise of the authority to enact the prohibition.

Section 9(c) is brief and unambiguous. It does not attempt to control the production of petroleum and petroleum products within a state. It does not seek to lay down rules for the guidance of state Legislatures or state officers. It leaves to the states and to their constituted authorities the determination of what production shall be permitted. It does not qualify the President’s authority by reference to the basis or extent of the state’s limitation of production. Section 9(c) does not state whether or in what circumstances or under what conditions the President is to prohibit the transportation of the amount of petroleum or petroleum products produced in excess of the state’s permission. It establishes no criterion to govern the President’s course. It does not require any finding by the President as a condition of his action. The Congress in § 9(c) thus declares no policy as to the transportation of the excess production. So far as this section is concerned, it gives to the President an unlimited authority to determine the policy and to lay down the prohibition, or not to lay it down, as he may see fit. And disobedience to his order is made a crime punishable by fine and imprisonment.

We examine the context to ascertain if it furnishes a declaration of policy or a standard of action, which can be deemed to relate to the subject of § 9(c) and thus to imply what is not there expressed. It is important to note that § 9 is headed “Oil Regulation”—that is, § 9 is the part of the National Industrial Recovery Act which particularly deals with that subject-matter. But the other provisions of § 9 afford no ground for implying a limitation of the broad grant of authority in § 9(c). . . .

We turn to the other provisions of Title I of the Act.
The first section is a “declaration of policy.” It declares that a national emergency exists which is “productive of widespread unemployment and disorganization of industry, which burdens interstate and foreign commerce, affects the public welfare, and undermines the standards of living of the American people.” It is declared to be the policy of Congress “to remove obstructions to the free flow of interstate and foreign commerce which tend to diminish the amount thereof”; “to provide for the general welfare by promoting the organization of industry for the purpose of cooperative action among trade groups”; “to induce and maintain united action of labor and management under adequate governmental sanctions and supervision”; “to eliminate unfair competitive practices, to promote the fullest possible utilization of the present productive capacity of industries, to avoid undue restriction of production (except as may be temporarily required), to increase the consumption of industrial and agricultural products by increasing purchasing power, to reduce and relieve unemployment, to improve standards of labor, and otherwise to rehabilitate industry and to conserve natural resources.”

This general outline of policy contains nothing as to the circumstances or conditions in which transportation of petroleum or petroleum products should be prohibited—nothing as to the policy of prohibiting or not prohibiting the transportation of production exceeding what the states allow. The general policy declared is “to remove obstructions to the free flow of interstate and foreign commerce.” As to production, the section lays down no policy of limitation. It favors the fullest possible utilization of the present productive capacity of industries. It speaks, parenthetically, of a possible temporary restriction of production, but of what, or in what circumstances, it gives no suggestion. The section also speaks in general terms of the conservation of natural resources, but it prescribes no policy for the achievement of that end. It is manifest that this broad outline is simply an introduction of the act, leaving the legislative policy as to particular subjects to be declared and defined, if at all, by the subsequent sections.

It is no answer to insist that deleterious consequences follow the transportation of “hot oil”—oil exceeding state allowances. The Congress did not prohibit that transportation. The Congress did not undertake to say that the transportation of “hot oil” was injurious. The Congress did not say that transportation of that oil was “unfair competition.” The Congress did not declare in what circumstances that transportation should be forbidden, or require the President to make any determination as to any facts or circumstances. Among the numerous and diverse objectives broadly stated, the President was not required to choose. The President was not required to ascertain and proclaim the conditions prevailing in the industry which made the prohibition necessary. The Congress left the matter to the President without standard or rule, to be dealt with as he pleased. The effort by ingenious and diligent construction to supply a criterion still permits such a breadth of authorized action as essentially to commit to the President the functions of a Legislature rather than those of an executive or administrative officer executing a declared legislative policy. We find nothing in § 1 which limits or controls the authority conferred by § 9(c). . . .

None of [the other sections of the act] can be deemed to prescribe any limitation of the grant of authority in § 9(c).
Fifth. The question whether such a delegation of legislative power is permitted by the Constitution is not answered by the argument that it should be assumed that the President has acted, and will act, for what he believes to be the public good. The point is not one of motives, but of constitutional authority, for which the best of motives is not a substitute. While the present controversy relates to a delegation to the President, the basic question has a much wider application. If the Congress can make a grant of legislative authority of the sort attempted by § 9(c), we find nothing in the Constitution which restricts the Congress to the selection of the President as grantee. The Congress may vest the power in the officer of its choice or in a board or commission such as it may select or create for the purpose. Nor, with respect to such a delegation, is the question concerned merely with the transportation of oil, or of oil produced in excess of what the state may allow. If legislative power may thus be vested in the President or other grantee as to that excess of production, we see no reason to doubt that it may similarly be vested with respect to the transportation of other commodities in interstate commerce with or without reference to state action, thus giving to the grantee of the power the determination of what is a wise policy as to that transportation, and authority to permit or prohibit it, as the person or board or commission so chosen may think desirable. In that view, there would appear to be no ground for denying a similar prerogative of delegation with respect to other subjects of legislation.

The Constitution provides that “All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.” Art. I, § 1. And the Congress is empowered “To make all Laws which shall be necessary and proper for carrying into Execution” its general powers. Art. I, § 8, par. 18. The Congress manifestly is not permitted to abdicate or to transfer to others the essential legislative functions with which it is thus vested. Undoubtedly legislation must often be adapted to complex conditions involving a host of details with which the national Legislature cannot deal directly. The Constitution has never been regarded as denying to the Congress the necessary resources of flexibility and practicality, which will enable it to perform its function in laying down policies and establishing standards, while leaving to selected instrumentalities the making of subordinate rules within prescribed limits and the determination of facts to which the policy as declared by the Legislature is to apply. Without capacity to give authorizations of that sort we should have the anomaly of a legislative power which in many circumstances calling for its exertion would be but a futility. But the constant recognition of the necessity and validity of such provisions and the wide range of administrative authority which has been developed by means of them cannot be allowed to obscure the limitations of the authority to delegate, if our constitutional system is to be maintained.

The Court has had frequent occasion to refer to these limitations and to review the course of congressional action. At the very outset, amid the disturbances due to war in Europe, when the national safety was imperiled and our neutrality was disregarded, the Congress passed a series of acts, as a part of which the President was authorized, in stated
circumstances, to lay and revoke embargoes, to give permits for the exportation of arms and military stores, to remit and discontinue the restraints and prohibitions imposed by acts suspending commercial intercourse with certain countries, and to permit or interdict the entrance into waters of the United States of armed vessels belonging to foreign nations. These early acts were not the subject of judicial decision, and, apart from that, they afford no adequate basis for a conclusion that the Congress assumed that it possessed an unqualified power of delegation. They were inspired by the vexations of American commerce through the hostile enterprises of the belligerent powers, they were directed to the effective execution of policies repeatedly declared by the Congress, and they confided to the President, for the purposes and under the conditions stated, an authority which was cognate to the conduct by him of the foreign relations of the government.

The first case relating to an authorization of this description was that of *The Brig Aurora* (1813). The cargo of that vessel had been condemned as having been imported from Great Britain in violation of the Nonintercourse Act of March 1, 1809. That act expired on May 1, 1810, when Congress passed another act providing that, in case either Great Britain or France before March 3, 1811, “shall . . . so revoke or modify her edicts as that they shall cease to violate the neutral commerce of the United States, which fact the President of the United States shall declare by proclamation, and if the other nation shall not within three months thereafter so revoke or modify her edicts in like manner,” then, with respect to that nation, as stated, the provisions of the act of 1809, after three months from that proclamation, “shall . . . be revived and have full force and effect.” On November 2, 1810, the President issued his proclamation declaring that France had so revoked or modified her edicts, and it was contended that the provisions of the act of 1809, as to the cargo in question, had thus been revived. The Court said that it could see no sufficient reason why the Legislature should not exercise its discretion in reviving the act of 1809, “either expressly or conditionally, as their judgment should direct.” The provision of that act declaring “that it should continue in force to a certain time, and no longer,” could not restrict the power of the Legislature to extend its operation “without limitation upon the occurrence of any subsequent combination of events.”

In *Field v. Clark* (1892), the Court applied that ruling to the case of “the suspension of an act upon a contingency to be ascertained by the president, and made known by his proclamation.” The Court was dealing with [a statute that] provided that, “with a view to secure reciprocal trade” with countries producing certain articles, “whenever, and so often as the President shall be satisfied” that the government of any country producing them imposed “duties or other exactions upon the agricultural or other products of the United States” which . . . the President “may deem to be reciprocally unequal and unreasonable, he shall have the power and it shall be his duty,” to suspend the free introduction of those articles by proclamation to that effect, and that during that suspension the duties specified by the section should be levied. The validity of the provision was challenged as a delegation to the President of legislative power. . . . While sustaining the provision, the Court emphatically declared that the principle that “congress cannot delegate legislative power to the president” is “universally recognized as vital to the integrity and maintenance of the system of government ordained by the constitution.” The Court found that the act before it was not inconsistent with that principle; that it did
not “in any real sense, invest the president with the power of legislation.” As “the
suspension was absolutely required when the president ascertained the existence of a
particular fact,” it could not be said “that in ascertaining that fact, and in issuing his
proclamation, in obedience to the legislative will, he exercised the function of making
laws.” “He was the mere agent of the law-making department to ascertain and declare the
event upon which its expressed will was to take effect.” The Court [distinguished]
between “the delegation of power to make the law, which necessarily involves a
discretion as to what it shall be, and conferring authority or discretion as to its execution,
to be exercised under and in pursuance of the law.”

Applying that principle, authorizations given by Congress to selected instrumentalities
for the purpose of ascertaining the existence of facts to which legislation is directed have
constantly been sustained. Moreover the Congress may not only give such authorizations
to determine specific facts, but may establish primary standards, devolving upon others
the duty to carry out the declared legislative policy; that is, as Chief Justice Marshall
expressed it, “to fill up the details” under the general provisions made by the Legislature.
In Butfield v. Stranahan, [a statute] was upheld, which authorized the Secretary of the
Treasury, upon the recommendation of a board of experts, to “establish uniform standards
of purity, quality, and fitness for consumption of all kinds of teas imported into the
United States.” The Court construed the statute as expressing “the purpose to exclude the
lowest grades of tea, whether demonstrably of inferior purity, or unfit for consumption, or
presumably so because of their inferior quality.” The Congress, the Court said, thus fixed
“a primary standard” and committed to the Secretary of the Treasury “the mere executive
duty to effectuate the legislative policy declared in the statute.” “Congress legislated on
the subject as far as was reasonably practicable, and from the necessities of the case was
compelled to leave to executive officials the duty of bringing about the result pointed out
by the statute.”

Another notable illustration is that of the authority given to the Secretary of War to
determine whether bridges and other structures constitute unreasonable obstructions to
navigation and to remove such obstructions. By that statute the Congress declared “a
general rule and imposed upon the Secretary of War the duty of ascertaining what
particular cases came within the rule” as thus laid down. Upon this principle rests the
authority of the Interstate Commerce Commission, in the execution of the declared policy
of the Congress in enforcing reasonable rates, in preventing undue preferences and unjust
discriminations, in requiring suitable facilities for transportation in interstate commerce,
and in exercising other powers held to have been validly conferred. Upon a similar
ground the authority given to the President, in appropriate relation to his functions as
Commander-in-Chief, by the Trading with the Enemy Act . . . with respect to the
disposition of enemy property, was sustained. “The determination,” said the Court, “of
the terms of sales of enemy properties in the light of facts and conditions from time to
time arising in the progress of war was not the making of a law; it was the application of
the general rule laid down by the act.”

The provisions of the Radio Act of 1927, providing for assignments of frequencies or
wave lengths to various stations, afford another instance. In granting licenses, the Radio
Commission is required to act “as public convenience, interest, or necessity requires.” In
construing this provision, the Court found that the statute itself declared the policy as to “equality of radio broadcasting service, both of transmission and of reception,” and that it conferred authority to make allocations and assignments in order to secure, according to stated criteria, an equitable adjustment in the distribution of facilities. The standard set up was not so indefinite “as to confer an unlimited power.”

So also, from the beginning of the Government, the Congress has conferred upon executive officers the power to make regulations,—“not for the government of their departments, but for administering the laws which did govern.” Such regulations become, indeed, binding rules of conduct, but they are valid only as subordinate rules and when found to be within the framework of the policy which the Legislature has sufficiently defined. In the case of United States v. Grimaud (1911), a regulation made by the Secretary of Agriculture requiring permits for grazing sheep on a forest reserve of lands belonging to the United States was involved. The Court referred to the various acts for the establishment and management of forest reservations and the authorization of rules which would “insure the objects of such reservations,” that is, “to regulate their occupancy and use and to preserve the forests thereon from destruction.” The Court observed that “it was impracticable for Congress to provide general regulations for these various and varying details of management,” and that, in authorizing the Secretary of Agriculture to meet local conditions, Congress “was merely conferring administrative functions upon an agent, and not delegating to him legislative power.” The Court quoted with approval the statement of the principle in Field, that the Congress cannot delegate legislative power, and upheld the regulation in question as an administrative rule for the appropriate execution of the policy laid down in the statute.

The applicable considerations were reviewed in Hampton & Co. v. United States (1928), where the Court dealt with the so-called “flexible tariff provision” of the Act of September 21, 1922, and with the authority which it conferred upon the President. The Court applied the same principle that permitted the Congress to exercise its rate-making power in interstate commerce, and found that a similar provision was justified for the fixing of customs duties; that is, as the Court said, “If Congress shall lay down by legislative act an intelligible principle to which the person or body authorized to fix such rates is directed to conform, such legislative action is not a forbidden delegation of legislative power. If it is thought wise to vary the customs duties according to changing conditions of production at home and abroad, it may authorize the Chief Executive to carry out this purpose, with the advisory assistance of a Tariff Commission appointed under congressional authority.” The Court sustained the provision upon the authority of Field, repeating with approval what was there said,—that “What the President was required to do was merely in execution of the act of Congress.”

Thus, in every case in which the question has been raised, the Court has recognized that there are limits of delegation which there is no constitutional authority to transcend. We think that § 9(c) goes beyond those limits. As to the transportation of oil production in excess of state permission, the Congress has declared no policy, has established no standard, has laid down no rule. There is no requirement, no definition of circumstances and conditions in which the transportation is to be allowed or prohibited.
If § 9(c) were held valid, it would be idle to pretend that anything would be left of limitations upon the power of the Congress to delegate its law-making function. The reasoning of the many decisions we have reviewed would be made vacuous and their distinctions nugatory. Instead of performing its law-making function, the Congress could at will and as to such subjects as it chooses transfer that function to the President or other officer or to an administrative body. The question is not of the intrinsic importance of the particular statute before us, but of the constitutional processes of legislation which are an essential part of our system of government.

Sixth. There is another objection to the validity of the prohibition laid down by the Executive Order under § 9(c). The Executive Order contains no finding, no statement of the grounds of the President’s action in enacting the prohibition. Both § 9(c) and the Executive Order are in notable contrast with historic practice . . . by which declarations of policy are made by the Congress and delegations are within the framework of that policy and have relation to facts and conditions to be found and stated by the President in the appropriate exercise of the delegated authority. If it could be said that from the four corners of the statute any possible inference could be drawn of particular circumstances or conditions which were to govern the exercise of the authority conferred, the President could not act validly without having regard to those circumstances and conditions. And findings by him as to the existence of the required basis of his action would be necessary to sustain that action, for otherwise the case would still be one of an unfettered discretion as the qualification of authority would be ineffectual. The point is pertinent in relation to the first section of the National Industrial Recovery Act. We have said that the first section is but a general introduction, that it declares no policy and defines no standard with respect to the transportation which is the subject of § 9(c). But if from the extremely broad description contained in that section and the widely different matters to which the section refers, it were possible to derive a statement of prerequisites to the President’s action under § 9(c), it would still be necessary for the President to comply with those conditions and to show that compliance as the ground of his prohibition. To hold that he is free to select as he chooses from the many and various objects generally described in the first section, and then to act without making any finding with respect to any object that he does select, and the circumstances properly related to that object, would be in effect to make the conditions inoperative and to invest him with an uncontrolled legislative power.

We are not dealing with action which, appropriately belonging to the executive province, is not the subject of judicial review or with the presumptions attaching to executive action. To repeat, we are concerned with the question of the delegation of legislative power. If the citizen is to be punished for the crime of violating a legislative order of an executive officer, or of a board or commission, due process of law requires that it shall appear that the order is within the authority of the officer, board, or commission, and, if that authority depends on determinations of fact, those determinations must be shown. As the Court said in *Wichita Railroad & Light Co. v. Public Utilities Comm’n*: “In creating such an administrative agency the legislature, to prevent its being a pure delegation of legislative power, must enjoin upon it a certain course of procedure and certain rules of decision in the performance of its function. It is a wholesome and necessary principle that such an agency must pursue the procedure and rules enjoined and show a substantial
compliance therewith to give validity to its action. When, therefore, such an administrative agency is required as a condition precedent to an order, to make a finding of facts, the validity of the order must rest upon the needed finding. If it is lacking, the order is ineffective. It is pressed on us that the lack of an express finding may be supplied by implication and by reference to the averments of the petition invoking the action of the Commission. We cannot agree to this.” . . . We cannot regard the President as immune from the application of these constitutional principles. When the President is invested with legislative authority as the delegate of Congress in carrying out a declared policy, he necessarily acts under the constitutional restriction applicable to such a delegation.

We see no escape from the conclusion that the Executive Orders . . . and . . . Regulations . . . are without constitutional authority.

**MR. JUSTICE CARDOZO, dissenting.**

. . . I am unable to assent to the conclusion that § 9(c) . . . is to be nullified upon the ground that [the President’s] discretion is too broad . . . . My point of difference with the majority of the court is narrow. I concede that to uphold the delegation there is need to discover in the terms of the act a standard reasonably clear whereby discretion must be governed. I deny that such a standard is lacking in respect of the prohibitions permitted by this section when the act with all its reasonable implications is considered as a whole. What the standard is becomes the pivotal inquiry.

As to the nature of the *act* which the President is authorized to perform there is no need for implication. That at least is definite beyond the possibility of challenge. He may prohibit the transportation in interstate and foreign commerce of petroleum and the products thereof produced or withdrawn from storage in excess of the amount permitted by any state law . . . . He is not left to roam at will among all the possible subjects of interstate transportation, picking and choosing as he pleases. I am far from asserting now that delegation would be valid if accompanied by all that latitude of choice. In the laying of his interdict he is to confine himself to a particular commodity, and to that commodity when produced or withdrawn from storage in contravention of the policy and statutes of the states. He has choice, though within limits, as to the occasion, but none whatever as to the means. The means have been prescribed by Congress. There has been no grant to the Executive of any roving commission to inquire into evils and then, upon discovering them, do anything he pleases. His act being thus defined, what else must he ascertain in order to regulate his discretion and bring the power into play? The answer is not given if we look to § 9(c) only, but it comes to us by implication from a view of other sections where the standards are defined. The prevailing opinion concedes that a standard will be as effective if imported into § 9(c) by reasonable implication as if put there in so many words. If we look to the whole structure of the statute, the test is plainly this, that the President is to forbid the transportation of the oil when he believes, in the light of the conditions of the industry as disclosed from time to time, that the prohibition will tend to effectuate the declared policies of the act,—not merely his own conception of its policies, undirected by any extrinsic guide, but the policies announced by § 1 in the forefront of the statute as an index to the meaning of everything that follows.
Oil produced or transported in excess of a statutory quota is known in the industry as “hot oil,” and the record is replete with evidence as to the effect of such production and transportation upon the economic situation and upon national recovery. A declared policy of Congress in the adoption of the act is “to eliminate unfair competitive practices.” Beyond question an unfair competitive practice exists when “hot oil” is transported in interstate commerce with the result that law-abiding dealers must compete with lawbreakers. Here is one of the standards set up in the act to guide the President’s discretion. Another declared policy of Congress is “to conserve natural resources.” Beyond question the disregard of statutory quotas is wasting the oil fields in Texas and other states and putting in jeopardy of exhaustion one of the treasures of the nation. All this is developed in the record and in the arguments of counsel for the government with a wealth of illustration. Here is a second standard. Another declared policy of Congress is to “promote the fullest possible utilization of the present productive capacity of industries,” and “except as may be temporarily required” to “avoid undue restriction of production.” Beyond question prevailing conditions in the oil industry have brought about the need for temporary restriction in order to promote in the long run the fullest productive capacity of business in all its many branches, for the effect of present practices is to diminish that capacity by demoralizing prices and thus increasing unemployment. The ascertainment of these facts at any time or place was a task too intricate and special to be performed by Congress itself through a general enactment in advance of the event. All that Congress could safely do was to declare the act to be done and the policies to be promoted, leaving to the delegate of its power the ascertainment of the shifting facts that would determine the relation between the doing of the act and the attainment of the stated ends. That is what it did. It said to the President in substance: You are to consider whether the transportation of oil in excess of the statutory quotas is offensive to one or more of the policies enumerated in § 1, whether the effect of such conduct is to promote unfair competition or to waste the natural resources or to demoralize prices or to increase unemployment or to reduce the purchasing power of the workers of the nation. If these standards or some of them have been flouted with the result of a substantial obstruction to industrial recovery, you may then by a prohibitory order eradicate the mischief.

I am not unmindful of the argument that the President has the privilege of choice between one standard and another, acting or failing to act according to an estimate of values that is individual and personal. To describe his conduct thus is to ignore the essence of his function. What he does is to inquire into the industrial facts as they exist from time to time. These being ascertained, he is not to prefer one standard to another in any subjective attitude of mind, in any personal or wilful way. He is to study the facts objectively, the violation of a standard impelling him to action or inaction according to its observed effect upon industrial recovery,—the ultimate end, as appears by the very heading of the title, to which all the other ends are tributary and mediate. Nor is there any essential conflict among the standards inter se, at all events when they are viewed in relation to § 9(c) and the power there conferred. In its immediacy, the exclusion of oil from the channels of transportation is a restriction of interstate commerce, not a removal of obstructions. This is self-evident, and, of course, was understood by Congress when the discretionary power of exclusion was given to its delegate. But what is restriction in its immediacy may in its ultimate and larger consequences be expansion and
development. Congress was aware that for the recovery of national well-being there might be need of temporary restriction upon production in one industry or another. It said so in § 1. When it clothed the President with power to impose such a restriction—to prohibit the flow of oil illegally produced—it laid upon him a mandate to inquire and determine whether the conditions in that particular industry were such at any given time as to make restriction helpful to the declared objectives of the act and to the ultimate attainment of industrial recovery. If such a situation does not present an instance of lawful delegation in a typical and classic form, categories long established will have to be formulated anew.

In what has been written, I have stated, but without developing the argument, that by reasonable implication the power conferred upon the President by § 9(c) is to be read as if coupled with the words that he shall exercise the power whenever satisfied that by doing so he will effectuate the policy of the statute as theretofore declared. Two canons of interpretation, each familiar to our law, leave no escape from that conclusion. One is that the meaning of a statute is to be looked for, not in any single section, but in all the parts together and in their relation to the end in view. The other is that, when a statute is reasonably susceptible of two interpretations, by one of which it is unconstitutional and by the other valid, the court prefers the meaning that preserves to the meaning that destroys. Plainly, § 1, with its declaration of the will of Congress, is the chart that has been furnished to the President to enable him to shape his course among the reefs and shallows of this act. If there could be doubt as to this when § 1 is viewed alone, the doubt would be dispelled by the reiteration of the policy in the sections that come later. In § 2, which relates to administrative agencies, in § 3, which relates to codes of Fair Competition, in § 4, which relates to agreements and licenses, in § 6, which prescribes limitations upon the application of the statute, and in § 10 which permits the adoption of rules and regulations, authority is conferred upon the President to do one or more acts as the delegate of Congress when he is satisfied that thereby he will aid “in effectuating the policy of this title” or in carrying out its provisions. True § 9, the one relating to petroleum, does not by express words of reference embody the same standard, yet nothing different can have been meant. What, indeed, is the alternative? Either the statute means that the President is to adhere to the declared policy of Congress, or it means that he is to exercise a merely arbitrary will. The one construction invigorates the act; the other saps its life. A choice between them is not hard.

I am persuaded that a reference, express or implied, to the policy of Congress as declared in § 1, is a sufficient definition of a standard to make the statute valid. Discretion is not unconfined and vagrant. It is canalized within banks that keep it from overflowing. . . . [T]he separation of powers between the Executive and Congress is not a doctrinaire concept to be made use of with pedantic rigor. There must be sensible approximation, there must be elasticity of adjustment, in response to the practical necessities of government, which cannot foresee today the developments of tomorrow in their nearly infinite variety. The Interstate Commerce Commission, probing the economic situation of the railroads of the country, consolidating them into systems, shaping in numberless ways their capacities and duties, and even making or unmaking the prosperity of great communities, is a conspicuous illustration. There could surely be no question as to the validity of an act whereby carriers would be prohibited from transporting oil produced in

Draft — do not distribute!
contravention of a statute if in the judgment of the Commission the practice was
demoralizing the market and bringing disorder and insecurity into the national economy.
What may be delegated to a commission may be delegated to the President. “Congress
may feel itself unable conveniently to determine exactly when its exercise of the
legislative power should become effective, because dependent on future conditions, and it
may leave the determination of such time to the decision of an executive.” . . . In the
complex life of today, the business of government could not go on without the delegation,
in greater or less degree, of the power to adapt the rule to the swiftly moving facts.

A striking illustration of this need is found in the very industry affected by this section,
the production of petroleum and its transportation between the states. At the passage of
the National Recovery Act no one could be certain how many of the states would adopt
valid quota laws, or how generally the laws would be observed when adopted, or to what
extent illegal practices would affect honest competitors or the stability of prices or the
conservation of natural resources or the return of industrial prosperity. Much would
depend upon conditions as they shaped themselves thereafter. Violations of the state laws
might turn out to be so infrequent that the honest competitor would suffer little, if any,
damage. The demand for oil might be so reduced that there would be no serious risk of
waste, depleting or imperiling the resources of the nation. Apart from these possibilities,
the business might become stabilized through voluntary cooperation or the adoption of a
code or otherwise. Congress not unnaturally was unwilling to attach to the state laws a
sanction so extreme as the cutting off of the privilege of interstate commerce unless the
need for such action had unmistakably developed. What was left to the President was to
ascertain the conditions prevailing in the industry, and prohibit or fail to prohibit
according to the effect of those conditions upon the phases of the national policy relevant
thereto. . . .

There is no fear that the nation will drift from its ancient moorings as the result of the
narrow delegation of power permitted by this section. What can be done under cover of
that permission is closely and clearly circumscribed both as to subject-matter and
occasion. The statute was framed in the shadow of a national disaster. A host of
unforeseen contingencies would have to be faced from day to day, and faced with a
fullness of understanding unattainable by any one except the man upon the scene. The
President was chosen to meet the instant need.

A subsidiary question remains as to the form of the executive order . . . . The question is a
subsidiary one, for, unless the statute is invalid, another order with fuller findings or
recitals may correct the informalities of this one, if informalities there are. But the order
to my thinking is valid as it stands. The President was not required either by the
Constitution or by any statute to state the reasons that had induced him to exercise the
granted power. It is enough that the grant of power had been made and that pursuant to
that grant he had signified the will to act. The will to act being declared, the law
presumes that the declaration was preceded by due inquiry and that it was rooted in
sufficient grounds. Such, for a hundred years and more, has been the doctrine of this
court. The act of February 28, 1795, authorized the President “whenever the United
States shall be invaded, or be in imminent danger of invasion from any foreign nation or
Indian tribe,” to call forth such number of the militia of the states as he shall deem
necessary and to issue his orders to the appropriate officers for that purpose. When war threatened in the summer of 1812, President Madison, acting under the authority of that statute, directed Major General Dearborn to requisition from New York, Massachusetts and Connecticut certain numbers of the states’ militia. No finding of “imminent danger of invasion” was made by the President in any express way, nor was such a finding made by the Secretary of War or any other official. . . . This does not mean that the individual is helpless in the face of usurpation. A court will not revise the discretion of the Executive, sitting in judgment on his order as if it were the verdict of a jury. On the other hand, we have said that his order may not stand if it is an act of mere oppression, an arbitrary fiat that overleaps the bounds of judgment. The complainants and others in their position may show, if they can, that in no conceivable aspect was there anything in the conditions of the oil industry in July, 1933, to establish a connection between the prohibitory order and the declared policies of the Congress. This is merely to say that the standard must be such as to have at least a possible relation to the act to be performed under the delegated power. One can hardly suppose that a prohibitory order would survive a test in court if the Executive were to assert a relation between the transportation of petroleum and the maintenance of the gold standard or the preservation of peace in Europe or the Orient. On the other hand, there can be no challenge of such a mandate unless the possibility of a rational nexus is lacking altogether. Here, in the case at hand, the relation between the order and the standard is manifest upon the face of the transaction from facts so notorious as to be within the range of our judicial notice. There is significance in the fact that it is not challenged even now.

The President, when acting in the exercise of a delegated power, is not a quasi-judicial officer, whose rulings are subject to review upon certiorari or appeal, or an administrative agency supervised in the same way. Officers and bodies such as those may be required by reviewing courts to express their decision in formal and explicit findings to the end that review may be intelligent. Such is not the position or duty of the President. He is the Chief Executive of the nation, exercising a power committed to him by Congress, and subject, in respect of the formal qualities of his acts, to the restrictions, if any, accompanying the grant, but not to any others. One will not find such restrictions either in the statute itself or in the Constitution back of it. The Constitution of the United States is not a code of civil practice. . . .

. . . If findings are necessary as a preamble to general regulations, the requirement must be looked for elsewhere than in the Constitution of the nation. . . .


MR. CHIEF JUSTICE HUGHES delivered the opinion of the Court.

. . . The “Live Poultry Code” was promulgated under § 3 of the National Industrial Recovery Act. That section . . . authorizes the President to approve “codes of fair
competition.” Such a code may be approved for a trade or industry, upon application by one or more trade or industrial associations or groups, if the President finds (1) that such associations or groups “impose no inequitable restrictions on admission to membership therein and are truly representative,” and (2) that such codes are not designed “to promote monopolies or to eliminate or oppress small enterprises and will not operate to discriminate against them, and will tend to effectuate the policy” of Title I of the Act. . . .

As a condition of his approval, the President may “impose such conditions (including requirements for the making of reports and the keeping of accounts) for the protection of consumers, competitors, employees, and others, and in furtherance of the public interest, and may provide such exceptions to and exemptions from the provisions of such code as the President in his discretion deems necessary to effectuate the policy herein declared.” Where such a code has not been approved, the President may prescribe one, either on his own motion or on complaint. . . .

The declared purpose [of the code] is “To effect the policies of title I of the National Industrial Recovery Act.” . . .

The code fixes [maximum hours and minimum wages, prohibits child labor, and contains union representation and other provisions].

Provision is made for administration through an “industry advisory committee,” to be selected by trade associations and members of the industry, and a “code supervisor” to be appointed, with the approval of the committee, by agreement between the Secretary of Agriculture and the Administrator for Industrial Recovery. . . .

The President approved the code by an executive order in which he found that the application for his approval had been duly made in accordance with the provisions of Title I of the National Industrial Recovery Act, that there had been due notice and hearings, that the Code constituted “a code of fair competition” as contemplated by the Act and complied with its pertinent provisions, including clauses (1) and (2) of subsection (a) of § 3 . . . ; and that the Code would tend “to effectuate the policy of Congress as declared in section 1” . . . .

[Defendants were convicted of violating various provisions of the code, including the minimum wage and maximum hour provisions.] . . .

. . . *The question of the delegation of legislative power.* . . . The Constitution provides that “All legislative powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.” And the Congress is authorized “To make all laws which shall be necessary and proper for carrying into execution” its general powers. The Congress is not permitted to abdicate or to transfer to others the essential legislative functions with which it is thus vested. We have repeatedly recognized the necessity of adapting legislation to complex conditions involving a host of details with which the national legislature cannot deal directly. We pointed out in the *Panama Company* case that the Constitution has never been regarded as denying to Congress the necessary resources of flexibility and practicality, which will enable it to perform its function in laying down policies and establishing standards, while leaving to
selected instrumentalities the making of subordinate rules within prescribed limits and the
determination of facts to which the policy as declared by the legislature is to apply. But
we said that the constant recognition of the necessity and validity of such provisions, and
the wide range of administrative authority which has been developed by means of them,
cannot be allowed to obscure the limitations of the authority to delegate, if our
constitutional system is to be maintained.

Accordingly, we look to the statute to see whether Congress has overstepped these
limitations,—whether Congress in authorizing “codes of fair competition” has itself
established the standards of legal obligation, thus performing its essential legislative
function, or, by the failure to enact such standards, has attempted to transfer that function
to others.

The aspect in which the question is now presented is distinct from that which was before
us in the case of the Panama Company. There the subject of the statutory prohibition was
defined. That subject was the transportation in interstate and foreign commerce of
petroleum and petroleum products which are produced or withdrawn from storage in
excess of the amount permitted by state authority. The question was with respect to the
range of discretion given to the President in prohibiting that transportation. As to the
“codes of fair competition,” under § 3 of the Act, the question is more fundamental. It is
whether there is any adequate definition of the subject to which the codes are to be
addressed.

What is meant by “fair competition” as the term is used in the Act? Does it refer to a
category established in the law, and is the authority to make codes limited accordingly?
Or is it used as a convenient designation for whatever set of laws the formulators of a
code for a particular trade or industry may propose and the President may approve
(subject to certain restrictions), or the President may himself prescribe, as being wise and
beneficient provisions for the government of the trade or industry in order to accomplish
the broad purposes of rehabilitation, correction, and expansion which are stated in [§ 1]?

The act does not define “fair competition.” “Unfair competition,” as known to the
common law, is a limited concept. Primarily, and strictly, it relates to the palming off of
one’s goods as those of a rival trader. In recent years, its scope has been extended. It has
been held to apply to misappropriation as well as misrepresentation, to the selling of
another’s goods as one’s own,—to misappropriation of what equitably belongs to a
competitor. Unfairness in competition has been predicated of acts which lie outside the
ordinary course of business and are tainted by fraud, or coercion, or conduct otherwise
prohibited by law. But it is evident that in its widest range, “unfair competition,” as it has
been understood in the law, does not reach the objectives of the codes which are
authorized by the National Industrial Recovery Act. The codes may, indeed, cover
conduct which existing law condemns, but they are not limited to conduct of that sort.
The Government does not contend that the Act contemplates such a limitation. It would
be opposed both to the declared purposes of the Act and to its administrative
construction.
The Federal Trade Commission Act . . . introduced the expression “unfair methods of competition,” which were declared to be unlawful. That was an expression new in the law. Debate apparently convinced the sponsors of the legislation that the words “unfair competition,” in the light of their meaning at common law, were too narrow. We have said that the substituted phrase has a broader meaning, that it does not admit of precise definition, its scope being left to judicial determination as controversies arise. What are “unfair methods of competition” are thus to be determined in particular instances, upon evidence, in the light of particular competitive conditions and of what is found to be a specific and substantial public interest. To make this possible, Congress set up a special procedure. A Commission, a quasi-judicial body, was created. Provision was made for formal complaint, for notice and hearing, for appropriate findings of fact supported by adequate evidence, and for judicial review to give assurance that the action of the Commission is taken within its statutory authority.

In providing for codes, the National Industrial Recovery Act dispenses with this administrative procedure and with any administrative procedure of an analogous character. But the difference between the code plan of the Recovery Act and the scheme of the Federal Trade Commission Act lies not only in procedure but in subject matter. We cannot regard the “fair competition” of the codes as antithetical to the “unfair methods of competition” of the Federal Trade Commission Act. The “fair competition” of the codes has a much broader range and a new significance. The Recovery Act provides that it shall not be construed to impair the powers of the Federal Trade Commission, but, when a code is approved, its provisions are to be the “standards of fair competition” for the trade or industry concerned, and any violation of such standards in any transaction in or affecting interstate or foreign commerce is to be deemed “an unfair method of competition” within the meaning of the Federal Trade Commission Act.

For a statement of the authorized objectives and content of the “codes of fair competition” we are referred repeatedly to the “Declaration of Policy” in section one . . . . Thus, the approval of a code by the President is conditioned on his finding that it “will tend to effectuate the policy of this title.” The President is authorized to impose such conditions “for the protection of consumers, competitors, employees, and others, and in furtherance of the public interest, and may provide such exceptions to and exemptions from the provisions of such code, as the President in his discretion deems necessary to effectuate the policy herein declared.” The “policy herein declared” is manifestly that set forth in section one. That declaration embraces a broad range of objectives. Among them we find the elimination of “unfair competitive practices.” But even if this clause were to be taken to relate to practices which fall under the ban of existing law, either common law or statute, it is still only one of the authorized aims described in section one. It is there declared to be “the policy of Congress”—

“to remove obstructions to the free flow of interstate and foreign commerce which tend to diminish the amount thereof; and to provide for the general welfare by promoting the organization of industry for the purpose of cooperative action among trade groups, to induce and maintain united action of labor and management under adequate governmental sanctions and supervision, to eliminate unfair competitive practices, to promote the fullest possible utilization of the
present productive capacity of industries, to avoid undue restriction of production (except as may be temporarily required), to increase the consumption of industrial and agricultural products by increasing purchasing power, to reduce and relieve unemployment, to improve standards of labor, and otherwise to rehabilitate industry and to conserve natural resources.”

Under § 3, whatever “may tend to effectuate” these general purposes may be included in the “codes of fair competition.” We think the conclusion is inescapable that the authority sought to be conferred by § 3 was not merely to deal with “unfair competitive practices” which offend against existing law, and could be the subject of judicial condemnation without further legislation, or to create administrative machinery for the application of established principles of law to particular instances of violation. Rather, the purpose is clearly disclosed to authorize new and controlling prohibitions through codes of laws which would embrace what the formulators would propose, and what the President would approve, prescribe, as wise and beneficent measures for the government of trades and industries in order to bring about their rehabilitation, correction and development, according to the general declaration of policy in section one. Codes of laws of this sort are styled “codes of fair competition.” . . .

The Government urges that the codes will “consist of rules of competition deemed fair for each industry by representative members of that industry—by the persons most vitally concerned and most familiar with its problems.” Instances are cited in which Congress has availed itself of such assistance; as, e.g., in the exercise of its authority over the public domain, with respect to the recognition of local customs or rules of miners as to mining claims, or, in matters of a more or less technical nature, as in designating the standard height of drawbars. But would it be seriously contended that Congress could delegate its legislative authority to trade or industrial associations or groups so as to empower them to enact the laws they deem to be wise and beneficent for the rehabilitation and expansion of their trade or industries? Could trade or industrial associations or groups be constituted legislative bodies for that purpose because such associations or groups are familiar with the problems of their enterprises? And could an effort of that sort be made valid by such a preface of generalities as to permissible aims as we find in section one . . . ? The answer is obvious. Such a delegation of legislative power is unknown to our law, and is utterly inconsistent with the constitutional prerogatives and duties of Congress.

The question, then, turns upon the authority which § 3 . . . vests in the President to approve or prescribe. If the codes have standing as penal statutes, this must be due to the effect of the executive action. But Congress cannot delegate legislative power to the President to exercise an unfettered discretion to make whatever laws he thinks may be needed or advisable for the rehabilitation and expansion of trade or industry.

Accordingly we turn to the Recovery Act to ascertain what limits have been set to the exercise of the President’s discretion. First, the President, as a condition of approval, is required to find that the trade or industrial associations or groups which propose a code, “impose no inequitable restrictions on admission to membership” and are “truly representative.” That condition, however, relates only to the status of the initiators of the
new laws and not to the permissible scope of such laws. Second, the President is required to find that the code is not “designed to promote monopolies or to eliminate or oppress small enterprises and will not operate to discriminate against them.” . . . But these restrictions leave virtually untouched the field of policy envisaged by section one, and, in that wide field of legislative possibilities, the proponents of a code, refraining from monopolistic designs, may roam at will, and the President may approve or disapprove their proposals as he may see fit. That is the precise effect of the further finding that the President is to make—that the code “will tend to effectuate the policy of this title.” While this is called a finding, it is really but a statement of an opinion as to the general effect upon the promotion of trade or industry of a scheme of laws. These are the only findings which Congress has made essential in order to put into operation a legislative code having the aims described in the “Declaration of Policy.”

Nor is the breadth of the President’s discretion left to the necessary implications of this limited requirement as to his findings. As already noted, the President in approving a code may impose his own conditions, adding to or taking from what is proposed, as “in his discretion” he thinks necessary “to effectuate the policy” declared by the Act. Of course, he has no less liberty when he prescribes a code on his own motion or on complaint, and he is free to prescribe one if a code has not been approved. . . . And this authority relates to a host of different trades and industries, thus extending the President’s discretion to all the varieties of laws which he may deem to be beneficial in dealing with the vast array of commercial and industrial activities throughout the country.

Such a sweeping delegation of legislative power finds no support in the decisions upon which the Government especially relies. By the Interstate Commerce Act, Congress has itself provided a code of laws regulating the activities of the common carriers subject to the Act, in order to assure the performance of their services upon just and reasonable terms, with adequate facilities and without unjust discrimination. Congress from time to time has elaborated its requirements, as needs have been disclosed. To facilitate the application of the standards prescribed by the Act, Congress has provided an expert body. That administrative agency, in dealing with particular cases, is required to act upon notice and hearing, and its orders must be supported by findings of fact which in turn are sustained by evidence. When the Commission is authorized to issue, for the construction, extension, or abandonment of lines, a certificate of “public convenience and necessity,” or to permit the acquisition by one carrier of the control of another, if that is found to be “in the public interest,” we have pointed out that these provisions are not left without standards to guide determination. The authority conferred has direct relation to the standards prescribed for the service of common carriers, and can be exercised only upon findings, based upon evidence, with respect to particular conditions of transportation.

Similarly, we have held that the Radio Act of 1927 established standards to govern radio communications and, in view of the limited number of available broadcasting frequencies, Congress authorized allocation and licenses. The Federal Radio Commission was created as the licensing authority, in order to secure a reasonable equality of opportunity in radio transmission and reception. The authority of the Commission to grant licenses “as public convenience, interest or necessity requires” was limited by the nature of radio communications, and by the scope, character, and quality of the services
to be rendered and the relative advantages to be derived through distribution of facilities. These standards established by Congress were to be enforced upon hearing, and evidence, by an administrative body acting under statutory restrictions adapted to the particular activity.

In Hampton & Co., the question related to the “flexible tariff provision” of the Tariff Act of 1922. We held that Congress had described its plan “to secure by law the imposition of customs duties on articles of imported merchandise which should equal the difference between the cost of producing in a foreign country the articles in question and laying them down for sale in the United States, and the cost of producing and selling like or similar articles in the United States.” As the differences in cost might vary from time to time, provision was made for the investigation and determination of these differences by the executive branch so as to make “the adjustments necessary to conform the duties to the standard underlying that policy and plan.” The Court found the same principle to be applicable in fixing customs duties as that which permitted Congress to exercise its rate-making power in interstate commerce, “by declaring the rule which shall prevail in the legislative fixing of rates” and then remitting “the fixing of such rates” in accordance with its provisions “to a rate-making body.” The Court fully recognized the limitations upon the delegation of legislative power.

To summarize and conclude upon this point: Section 3 of the Recovery Act is without precedent. It supplies no standards for any trade, industry, or activity. It does not undertake to prescribe rules of conduct to be applied to particular states of fact determined by appropriate administrative procedure. Instead of prescribing rules of conduct, it authorizes the making of codes to prescribe them. For that legislative undertaking, § 3 sets up no standards, aside from the statement of the general aims of rehabilitation, correction and expansion described in section one. In view of the scope of that broad declaration and of the nature of the few restrictions that are imposed, the discretion of the President in approving or prescribing codes, and thus enacting laws for the government of trade and industry throughout the country, is virtually unfettered. We think that the code-making authority thus conferred is an unconstitutional delegation of legislative power.

MR. JUSTICE CARDozo, concurring.

The delegated power of legislation which has found expression in this code is not canalized within banks that keep it from overflowing. It is unconfined and vagrant, if I may borrow my own words in an earlier opinion. Panama Refining.

This court has held that delegation may be unlawful, though the act to be performed is definite and single, if the necessity, time, and occasion of performance have been left in the end to the discretion of the delegate. I thought that ruling went too far. I pointed out in an opinion that there had been “no grant to the Executive of any roving commission to inquire into evils and then, upon discovering them, do anything he pleases.” Choice, though within limits, had been given him “as to the occasion, but none whatever as to the means.” Here, in the case before us, is an attempted delegation not confined to any single act nor to any class or group of acts identified or described by reference to a standard.
Here in effect is a roving commission to inquire into evils and upon discovery correct them.

I have said that there is no standard, definite or even approximate, to which legislation must conform. Let me make my meaning more precise. If codes of fair competition are codes eliminating “unfair” methods of competition ascertained upon inquiry to prevail in one industry or another, there is no unlawful delegation of legislative functions when the President is directed to inquire into such practices and denounce them when discovered. For many years a like power has been committed to the Federal Trade Commission with the approval of this court in a long series of decisions. Delegation in such circumstances is born of the necessities of the occasion. The industries of the country are too many and diverse to make it possible for Congress, in respect of matters such as these, to legislate directly with adequate appreciation of varying conditions. Nor is the substance of the power changed because the President may act at the instance of trade or industrial associations having special knowledge of the facts. Their function is strictly advisory; it is the imprimatur of the President that begets the quality of law. When the task that is set before one is that of cleaning house, it is prudent as well as usual to take counsel of the dwellers.

But there is another conception of codes of fair competition, their significance and function, which leads to very different consequences, though it is one that is struggling now for recognition and acceptance. By this other conception a code is not to be restricted to the elimination of business practices that would be characterized by general acception as oppressive or unfair. It is to include whatever ordinances may be desirable or helpful for the well-being or prosperity of the industry affected. In that view, the function of its adoption is not merely negative, but positive; the planning of improvements as well as the extirpation of abuses. What is fair, as thus conceived, is not something to be contrasted with what is unfair or fraudulent or tricky. The extension becomes as wide as the field of industrial regulation. If that conception shall prevail, anything that Congress may do within the limits of the commerce clause for the betterment of business may be done by the President upon the recommendation of a trade association by calling it a code. This is delegation running riot. No such plenitude of power is susceptible of transfer. The statute, however, aims at nothing less, as one can learn both from its terms and from the administrative practice under it. Nothing less is aimed at by the code now submitted to our scrutiny.

The code does not confine itself to the suppression of methods of competition that would be classified as unfair according to accepted business standards or accepted norms of ethics. It sets up a comprehensive body of rules to promote the welfare of the industry, if not the welfare of the nation, without reference to standards, ethical or commercial, that could be known or predicted in advance of its adoption. One of the new rules . . . is aimed at an established practice, not unethical or oppressive, the practice of selective buying. Many others could be instanced as open to the same objection if the sections of the code were to be examined one by one. . . . Even if the statute itself had fixed the meaning of fair competition by way of contrast with practices that are oppressive or unfair, the code outruns the bounds of the authority conferred. What is excessive is not sporadic or superficial. It is deep-seated and pervasive. . . .

MR. JUSTICE SUTHERLAND delivered the opinion of the Court.

The purposes of the “Bituminous Coal Conservation Act of 1935” . . . are to stabilize the bituminous coal-mining industry and promote its interstate commerce; to provide for coöperative marketing of bituminous coal; . . . to declare the production, distribution, and use of such coal to be affected with a national public interest; to conserve the national resources of such coal; to provide for the general welfare . . . .

[The Act] establishes in the Department of the Interior a National Bituminous Coal Commission . . . [, which is required to] formulate . . . a working agreement to be known as the Bituminous Coal Code. [The Act] require[s] the organization of twenty-three coal districts, each with a district board the membership of which is to be determined in a manner pointed out by the act . . .

[The “Labor Relations” subdivision] provides:

“Whenever the maximum daily and weekly hours of labor are agreed upon in any . . . contracts negotiated between the producers of more than two-thirds the annual national tonnage production . . . and the representatives of more than one-half the mine workers employed, such maximum hours of labor shall be accepted by all the code members. . . . [W]age . . . agreements negotiated by collective bargaining [in individual districts] . . . shall be accepted as the minimum wages . . . by the code members operating in such . . . districts.”

. . . That the act, whatever it may be in form, in fact is compulsory clearly appears. [Section 3 imposes a 15% tax on the sale of all bituminous coal, but gives a 90% tax credit to any coal producer that accepts the code; this tax is] a penalty to compel “acceptance” of the code. Section 14 provides that the United States shall purchase no . . . coal [from any non-code producer] . . . . In the light of these provisions we come to a consideration of [the] subdivision . . . dealing with “Labor Relations.”

That subdivision delegates the power to fix maximum hours of labor to a part of the producers and the miners—namely, “the producers of more than two-thirds the annual national tonnage production for the preceding calendar year” and “more than one-half the mine workers employed”; and to producers of more than two-thirds of the district annual tonnage during the preceding calendar year and a majority of the miners, there is delegated the power to fix minimum wages . . . . The effect, in respect of wages and hours, is to subject the dissentient minority, either of producers or miners or both, to the will of the stated majority, since, by refusing to submit, the minority at once incurs the hazard of enforcement of the drastic compulsory provisions of the act to which we have referred. To “accept,” in these circumstances, is not to exercise a choice, but to surrender to force.
The power conferred upon the majority is, in effect, the power to regulate the affairs of an unwilling minority. This is legislative delegation in its most obnoxious form; for it is not even delegation to an official or an official body, presumptively disinterested, but to private persons whose interests may be and often are adverse to the interests of others in the same business. The record shows that the conditions of competition differ among the various localities. In some, coal dealers compete among themselves. In other localities, they also compete with the mechanical production of electrical energy and of natural gas. Some coal producers favor the code; others oppose it; and the record clearly indicates that this diversity of view arises from their conflicting and even antagonistic interests. The difference between producing coal and regulating its production is, of course, fundamental. The former is a private activity; the latter is necessarily a governmental function, since, in the very nature of things, one person may not be intrusted with the power to regulate the business of another, and especially of a competitor. And a statute which attempts to confer such power undertakes an intolerable and unconstitutional interference with personal liberty and private property. The delegation is so clearly arbitrary, and so clearly a denial of rights safeguarded by the due process clause of the Fifth Amendment, that it is unnecessary to do more than refer to decisions of this court which foreclose the question. [Cites Schechter Poultry and other cases].

Separate opinion of MR. CHIEF JUSTICE HUGHES.

. . . I also agree that [the Labor Relations] subdivision . . . is invalid . . . :

(1) It attempts a broad delegation of legislative power to fix hours and wages without standards of limitation. The Government invokes the analogy of legislation which becomes effective on the happening of a specified event, and says that in this case the event is the agreement of a certain proportion of producers and employees, whereupon the other producers and employees become subject to legal obligations accordingly. I think that the argument is unsound and is pressed to the point where the principle would be entirely destroyed. It would remove all restrictions upon the delegation of legislative power, as the making of laws could thus be referred to any designated officials or private persons whose orders or agreements would be treated as “events,” with the result that they would be invested with the force of law having penal sanctions.

(2) The provision permits a group of producers and employees, according to their own views of expediency, to make rules as to hours and wages for other producers and employees who were not parties to the agreement. Such a provision, apart from the mere question of the delegation of legislative power, is not in accord with the requirement of due process of law which under the Fifth Amendment dominates the regulations which Congress may impose. . . .

[Justice Cardozo wrote a separate opinion, in which he argued that the challenge to the Labor Relations section was premature.]
§ 1.5. Notes on the earlier federal cases

1. The evolution of the non-delegation doctrine. The *Panama Refining* opinion recapitulates the stages of the development of non-delegation doctrine. The inquiry has always been “whether the Congress has declared a policy [or] set up a standard for the [delegate’s] action.” The early cases, for instance *The Brig Aurora* and *Field*, rest on the “contingency rationale”—all the President needs to do is determine whether a particular fact exists. But, as *Field* makes clear, this is not a mechanistic factfinding exercise—can the “unreasonable[ness]” of foreign tariffs plausibly be described as a mere fact? Later cases, like *Buttfield* and *Grimaud*, described the delegate’s power as the power to “fill up the details,” for instance to establish tea standards or grazing regulations protective of forests. But, in the context of the Radio Act of 1927, defining “public convenience, interest, or necessity” is hardly a detail-oriented exercise. *Hampton* settled on the “intelligible principle” test, which is still used today. But see Justice Thomas’s dissent in *American Trucking*, infra. Later cases put more weight on the existence of external checks. See note 3 infra.

2. The role of statutory interpretation and the use of “nondelegation canons.”

Statutory interpretation plays two roles in these cases.

First, finding “a standard for the President’s action” requires reading the statute—first the text of the disputed section itself (the paragraph on § 9(c) in *Panama Refining*), next the statutory context (the discussion of the policies listed in § 1 and the other sections). The other tools of statutory interpretation can also come into play—one can also read the statute “purposively” (i.e., with reference to its broad purpose, “national recovery”) or “intentionally” (i.e., with reference to the intent of the legislators, as revealed by congressional committee reports or legislators’ statements during floor debate). In *Panama Refining*, Cardozo refers to contextualism and purposivism as a “canon” (i.e., rule or principle) of statutory interpretation: “the meaning of a statute is to be looked for, not in any single section, but in all the parts together and in their relation to the end in view.” The difference in opinion between the majority and Cardozo can be interpreted as a difference in how purposively to read the statute; the majority sees in § 1 a mishmash of competing policies, while Cardozo reads them as subservient to the overall goal of national recovery. For examples of intentionalism, see the opinions in the Benzene case, infra, which refer to “legislative history” (though the actual discussions of legislative history are not excerpted here).

The second role of statutory interpretation is in the use of canons of statutory interpretation to save a statute from unconstitutionality. Cardozo’s second canon is a case in point: “when a statute is reasonably susceptible of two interpretations, by one of which it is unconstitutional and by the other valid, the court prefers the meaning that preserves to the meaning that destroys.” This canon serves the function of “constitutional avoidance”: Usually, when an ambiguous statute has two available meanings, we might want to choose whichever one is even slightly
more plausible. But suppose one of the meanings looks like it might be so broad as to produce an unconstitutional delegation. Then we (as judges) can invoke the canon and choose the narrower meaning, even if it’s not the one we would have chosen otherwise.

This approach avoids the unattractive (to some) option of overruling the political branches, and in fact allows Congress to reenact the more delegating version of the statute (without the ambiguity) if it so desires, which is why it is also called a “clear statement” rule. For another example of the same canon, see the majority opinion in the Benzene case, infra; another is Kent v. Dulles (1958). In Cass Sunstein’s view, even though the non-delegation doctrine is never used today to invalidate statutes outright, it hasn’t died but has only morphed into an avoidance canon. Sunstein, Nondelegation Canons, 67 U. CHI. L. REV. 315 (2000). The doctrine thus has real effects, even though it hasn’t invalidated a statute since the 1930s. As Cardozo says in Panama Refining, “[a] choice between [a constitutional reading and an unconstitutional one] is not hard.”

But not everyone agrees with Cardozo: the avoidance canon doesn’t require a court to find that a delegation would be unconstitutional, only that it might be. Therefore, courts, in effect, rewrite statutes based on their fears of what might be unconstitutional—resulting in a statute that may not be the “best” reading of the statutory text and that may not correspond with legislators’ intention when enacting the statute—even when those fears are ill-founded. See, e.g., Posner, The Federal Courts: Crisis and Reform (1985).

3. Non-delegation or due process? Part I—the importance of findings or procedures. In part “Sixth” of Panama Refining, the majority invalidates the President’s action in part because his Executive Order was not accompanied by any “finding” or “statement of the grounds of the President’s action in enacting the prohibition.” Note that this is not a non-delegation issue. First, the absence of findings doesn’t invalidate the statute, but would invalidate this particular action of the President’s even if one could find an intelligible principle (“if . . . it were possible to derive a statement of prerequisites to the President’s action under § 9(c)). Second, the majority explicitly locates the source of the requirements of findings in “due process of law,” that is, in the Fifth Amendment, not the Vesting Clause (Art. I, § 1).

But in Schechter Poultry, the existence of procedural safeguards is brought into the non-delegation analysis. For instance, the FTC Act is not an overbroad delegation, in part because the delegate is a “quasi-judicial body,” “notice and hearing,” and “judicial review”; the Interstate Commerce Act likewise must “act upon notice and hearing” and produce “findings of fact . . . sustained by evidence”; the Radio Act standards are “enforced upon hearing, and evidence”; and so on. This approach is also common in the more modern cases, for instance Yakus v. United States (1944) and Fahey v. Mallonee (1947).

On the one hand, a delegation with procedural safeguards is narrower than one
implemented without them, since otherwise the delegate might get away with more unauthorized acts and invalid exercises of his discretion. On the other hand, doesn’t this merge the non-delegation inquiry (Has Congress given up too much power?) and the due process inquiry (Has the delegate acted within the bounds of his power?)? See Posner & Vermeule, *Interring the Nondelegation Doctrine*, 69 U. CHI. L. REV. 1721 (2002).

4. **Non-delegation or due process? Part II—the puzzle of *Carter Coal*.** In *Schechter Poultry*, the Court was concerned about the private delegation (“Such a delegation of legislative power is unknown to our law”), but its decision rested on the absence of standards to constrain the President, who has the ultimately authority to approve the codes. *Carter Coal*, where private groups have the final say on labor relations, has a stronger condemnation of the perils of private delegation (“the power to regulate the affairs of an unwilling minority . . . is legislative delegation in its most obnoxious form”). But, for all its discussion of “delegation” and citation of *Schechter Poultry*, the basic concern is with having a biased decisionmaker (“private persons whose interests may be . . . adverse to the interests of others in the same business”), which is a traditional concern of due process analysis. Moreover, the holding is that the procedure involves “an intolerable and unconstitutional interference with personal liberty and private property,” is “arbitrary,” and denies “rights safeguarded by the due process clause.” Therefore, it seems to make more sense to view *Carter Coal* as a due process case, not a non-delegation case at all. See n.5 of the *Tex. Boll Weevil* dissent; *Mistretta v. United States* (1989).

In a way, perhaps this makes sense. If one is to construct a specific argument against private delegations, due process analysis seems inherently better suited to the task than non-delegation analysis. As noted above, the non-delegation inquiry seems more accurately focused on the extent of the power Congress has given up—which seems independent of the identity of the delegate—while the due process inquiry seems more appropriate for focusing on whether the delegate is using its delegated power constitutionally, taking into account notions of bias and arbitrariness. David Lawrence suggests that state courts, for private delegations, replace non-delegation analysis with due process analysis. Lawrence, *Private Exercise of Governmental Power*, 61 IND. L.J. 647 (1986); see also Jaffe, *Law Making by Private Groups*, 51 HARV. L. REV. 201 (1937). State private delegations are often analyzed under the due process clause. See also Abramson, *A Fifth Branch of Government: The Private Regulators and Their Constitutionality*, 16 HASTINGS Const. L.Q. 165 (1989); *Wash. ex rel. Seattle Title Trust Co. v. Roberge* (1928); *New Motor Veh. Bd. v. Orrin W. Fox Co.* (1978) (discussed in note 9 below).

5. **Public-private essentialism.** It is hard to argue with the *Carter Coal* majority that delegating the conduct of an industry to members of that industry raises the danger of bias and conflicts of interest. Do you agree that, on the other hand, “delegation to an official or an official body [is] presumptively disinterested”? The majority says that “in the very nature of things, one person may not be
intrusted with the power to regulate the business of another.” Who do you think staffs official bodies—persons, or something else? If there is a difference between the degree of self-interested behavior in official bodies and industry groups, is that difference a matter of degree, to be determined empirically, or is it an sharp difference that inheres “in the very nature of things”? This sort of public-private essentialism is a common theme in arguments that seek to distinguish public and private delegations. See also Tex. Boll Weevil, infra.

6. Private delegations and policing fuzzy boundaries. If, as the early cases show, policing delegations is a question of degree, if everyone agrees that some delegations are unavoidable and beneficial, and if courts are generally reluctant to second-guess legislative policy choices, is it any surprise that the non-delegation doctrine is difficult (or, depending on one’s view, impossible) to enforce? In Mistretta, Justice Scalia explained that, because of courts’ limitations, non-delegation is an underenforced constitutional norm—which makes it all the more important to pay attention to structural restrictions that would deter excessive delegation. Scalia was referring to the quasi-legislative power of the Sentencing Commission, but one can make the same argument about private delegations. If one believes that, overall, private delegations will raise more severe non-delegation concerns, but that those delegations will be impossible to police on a case-by-case basis, one can make out a systemic case for opposing private delegations as a class.

7. Private industry holding the on-off switch: Currin v. Wallace, 306 U.S. 1 (1939). The Tobacco Inspection Act regulated tobacco auction markets. The Secretary of Agriculture was authorized to designate tobacco markets that were to be governed by inspection and certification standards. But no market was to be designated unless this was requested by a two-thirds vote in a referendum of the participating growers. Recall the President’s limited power in Panama Refining, who, as Cardozo put it, “ha[d] choice, though within limits, as to the occasion, but none whatever as to the means.” In effect, once a state banned the production or withdrawal of oil, the President held an “on-off switch” as to the legality of its transportation under federal law. Nonetheless, the Court held this to be an invalid delegation because the President was given no standards to guide his discretion as to whether to flip the switch. Here, tobacco growers were given equally unconstrained discretion as to whether to flip the on-off switch of the Secretary’s potential designation of their market. Would a Panama Refining analysis similarly find this power invalid—perhaps even more so because it might be informed by a Carter Coal-style concern over the self-interestedness of industry participants?

The Court upheld the delegation. Chief Justice Hughes wrote: “[T]he required referendum does not involve any delegation of legislative authority [to tobacco growers]. Congress has merely placed a restriction upon its own regulation by withholding its operation as to a given market ‘unless two-thirds of the growers voting favor it.’ . . . This is not a case where a group of producers may make the law and force it upon a minority (see Carter Coal) or where a prohibition of an inoffensive and legitimate use of property is imposed not by the legislature but by
other property owners. Here it is Congress that exercises its legislative authority in making the regulation and in prescribing the conditions of its application. The required favorable vote upon the referendum is one of these conditions.

“The distinction was pointed out in *Hampton & Co. v. United States*, where, in sustaining the so-called ‘flexible tariff provision’ . . . and the authority it conferred upon the President, we said: ‘Congress may feel itself unable conveniently to determine exactly when its exercise of the legislative power should become effective, because dependent on future conditions, and it may leave the determination of such time to the decision of an Executive, or . . . a popular vote of the residents of a district to be affected by the legislation. While in a sense one may say that such residents are exercising legislative power, it is not an exact statement, because the power has already been exercised legislatively by [Congress], the condition of its legislation going into effect being made dependent . . . on the expression of the voters of a certain district.’”

Admittedly, the on-off switch power held by the industry is less than the broad regulatory power over labor relations held by the industry in *Carter Coal*. Nonetheless, can this be squared with *Panama Refining*?

8. **No special doctrine for private delegations.** Note that, in analogizing the industry’s power to the President’s in *Hampton & Co.*, Chief Justice Hughes in *Currin* confirmed that the test for invalid private delegations is the same as the test for invalid delegations generally. See also *Thomas v. Union Carbide Agric. Prods. Co.* (1985).

9. **Currin, state delegations, and due process: New Motor Vehicle Bd. v. Orrin W. Fox Co., 439 U.S. 96 (1978).** California’s Automobile Franchise Act required potential auto manufacturers to obtain the approval of the New Motor Vehicle Board before opening a dealership within 10 miles of one of their existing franchisees—but only if the existing franchisee complained. The Supreme Court, in an opinion by Justice Brennan, held that this was not an “impermissible delegation of state power to private citizens. . . . Almost any system of private or quasi-private law could be subject to the same objection. Court approval of an eviction, for example, becomes necessary only when the tenant protests his eviction, and he alone decides whether he will protest. An otherwise valid regulation is not rendered invalid simply because those whom the regulation is designed to safeguard may elect to forgo its protection.”

Note that, though the case said the word “delegation,” this was a due process opinion, since non-delegation doctrine in the classic sense derives from the Vesting Clause and therefore only applies to federal delegations (see note 1 of the General notes on the non-delegation doctrine); so *Currin* is not directly on point. Justice Stevens’s dissent analogized the process—where a private party could invoke a harmful government process, in this case the temporary delay associated with a Board hearing, without any government oversight—to the procedure
invalidated in *Fuentes v. Shevin* (1972), discussed below in the section on the State Action Doctrine.

10. **Private industry as mere advisors:** *Sunshine Anthracite Coal Co. v. Adkins*, 310 U.S. 381 (1940). After *Carter Coal*, Congress enacted a new Bituminous Coal Act. Coal producers were to be organized into a Bituminous Coal Code (with inducements as strong as in *Carter Coal*). “Some twenty district boards of code members are provided for, which are to operate as an aid to the Commission but subject to its pervasive surveillance and authority.” The government’s National Bituminous Coal Commission was empowered to fix minimum prices for code members after each district board “on its own motion or when directed by the Commission’ propose[d] minimum prices pursuant to prescribed statutory standards. These may be approved, disapproved, or modified by the Commission as the basis for the coördination of minimum prices.”

The Court upheld the statute. Justice Douglas wrote: “Nor has Congress delegated its legislative authority to the industry. The members of the code function subordinately to the Commission. It, not the code authorities, determines the prices. And it has authority and surveillance over the activities of these authorities. Since law-making is not entrusted to the industry, this statutory scheme is unquestionably valid.”

Note that in *Sunshine Anthracite*—and still in current law—government approval of a privately proposed scheme is all that’s required to validate the private delegation, or, more properly, to find that no private delegation has occurred at all. Whether the government submits the private proposal to searching review, or no review at all, is immaterial. When such a scheme is proposed, which do you think is more likely, searching review or rubber-stamping? In theory, the government official who rubber-stamps the private proposal can be held politically accountable. Will this also be true in practice? Should this make a difference as a matter of delegation doctrine?

---


**MR. JUSTICE STEVENS announced the judgment of the Court. . . .**

. . . This litigation concerns a standard promulgated by the Secretary of Labor [under the Occupational Safety and Health Act] to regulate occupational exposure to benzene, a substance which has been shown to cause cancer at high exposure levels. . . .

---

* Justice Stevens’ plurality opinion was joined, in the part quoted here, by Chief Justice Burger and Justices Stewart and Powell.
The Act delegates broad authority to the Secretary to promulgate different kinds of standards. The basic definition of an “occupational safety and health standard” is found in § 3(8), which provides:

“The term ‘occupational safety and health standard’ means a standard which requires conditions, or the adoption or use of one or more practices, means, methods, operations, or processes, reasonably necessary or appropriate to provide safe or healthful employment and places of employment.”

Where toxic materials or harmful physical agents are concerned, a standard must also comply with § 6(b)(5), which provides:

“The Secretary, in promulgating standards dealing with toxic materials or harmful physical agents under this subsection, shall set the standard which most adequately assures, to the extent feasible, on the basis of the best available evidence, that no employee will suffer material impairment of health or functional capacity even if such employee has regular exposure to the hazard dealt with by such standard for the period of his working life. Development of standards under this subsection shall be based upon research, demonstrations, experiments, and such other information as may be appropriate. In addition to the attainment of the highest degree of health and safety protection for the employee, other considerations shall be the latest available scientific data in the field, the feasibility of the standards, and experience gained under this and other health and safety laws.”

... [Originally, OSHA adopted a 10 ppm standard. Later, it adopted a more stringent 1 ppm standard. In its published statement giving notice of [this new proposed] standard, OSHA did not ask for comments as to whether or not benzene presented a significant health risk at exposures of 10 ppm or less. Rather, it asked for comments as to whether 1 ppm was the minimum feasible exposure limit. ... [T]his formulation of the issue ... was consistent with OSHA’s general policy with respect to carcinogens. Whenever a carcinogen is involved, OSHA will presume that no safe level of exposure exists in the absence of clear proof establishing such a level and will accordingly set the exposure limit at the lowest level feasible. The proposed 1 ppm exposure limit in this case thus was established not on the basis of a proven hazard at 10 ppm, but rather on the basis of “OSHA’s best judgement at the time of the proposal of the feasibility of compliance with the proposed standard by the [a]ffected industries.” Given OSHA’s cancer policy, it was in fact irrelevant whether there was any evidence at all of a leukemia risk at 10 ppm. The important point was that there was no evidence that there was not some risk, however small, at that level. The fact that OSHA did not ask for comments on whether there was a safe level of exposure for benzene was indicative of its further view that a demonstration of such absolute safety simply could not be made. ...
for toxic materials to “set the standard which most adequately assures, to the extent feasible, on the basis of the best available evidence, that no employee will suffer material impairment of health or functional capacity.” . . .

Under the Government’s view, § 3(8), if it has any substantive content at all, merely requires OSHA to issue standards that are reasonably calculated to produce a safer or more healthy work environment. Apart from this minimal requirement of rationality, the Government argues that § 3(8) imposes no limits on the Agency’s power, and thus would not prevent it from requiring employers to do whatever would be “reasonably necessary” to eliminate all risks of any harm from their workplaces. With respect to toxic substances and harmful physical agents, the Government takes an even more extreme position. Relying on § 6(b)(5)’s direction to set a standard “which most adequately assures . . . that no employee will suffer material impairment of health or functional capacity,” the Government contends that the Secretary is required to impose standards that either guarantee workplaces that are free from any risk of material health impairment, however small, or that come as close as possible to doing so without ruining entire industries.

If the purpose of the statute were to eliminate completely and with absolute certainty any risk of serious harm, we would agree that it would be proper for the Secretary to interpret §§ 3(8) and 6(b)(5) in this fashion. But we think it is clear that the statute was not designed to require employers to provide absolutely risk-free workplaces whenever it is technologically feasible to do so, so long as the cost is not great enough to destroy an entire industry. Rather, both the language and structure of the Act, as well as its legislative history, indicate that it was intended to require the elimination, as far as feasible, of significant risks of harm.

By empowering the Secretary to promulgate standards that are “reasonably necessary or appropriate to provide safe or healthful employment and places of employment,” the Act implies that, before promulgating any standard, the Secretary must make a finding that the workplaces in question are not safe. But “safe” is not the equivalent of “risk-free.” There are many activities that we engage in every day—such as driving a car or even breathing city air—that entail some risk of accident or material health impairment; nevertheless, few people would consider these activities “unsafe.” Similarly, a workplace can hardly be considered “unsafe” unless it threatens the workers with a significant risk of harm.

Therefore, before he can promulgate any permanent health or safety standard, the Secretary is required to make a threshold finding that a place of employment is unsafe—in the sense that significant risks are present and can be eliminated or lessened by a change in practices. This requirement applies to permanent standards promulgated pursuant to § 6(b)(5), as well as to other types of permanent standards. For there is no reason why § 3(8)’s definition of a standard should not be deemed incorporated by reference into § 6(b)(5). The standards promulgated pursuant to § 6(b)(5) are just one species of the genus of standards governed by the basic requirement. That section repeatedly uses the term “standard” without suggesting any exception from, or qualification of, the general definition; on the contrary, its directs the Secretary to select “the standard”—that is to say, one of various possible alternatives that satisfy the basic
definition in § 3(8)—that is most protective. Moreover, requiring the Secretary to make a threshold finding of significant risk is consistent with the scope of the regulatory power granted to him by § 6(b)(5), which empowers the Secretary to promulgate standards, not for chemicals and physical agents generally, but for “toxic materials” and “harmful physical agents” . . . .

In the absence of a clear mandate in the Act, it is unreasonable to assume that Congress intended to give the Secretary the unprecedented power over American industry that would result from the Government’s view of §§ 3(8) and 6(b)(5), coupled with OSHA’s cancer policy. Expert testimony that a substance is probably a human carcinogen—either because it has caused cancer in animals or because individuals have contracted cancer following extremely high exposures—would justify the conclusion that the substance poses some risk of serious harm no matter how minute the exposure and no matter how many experts testified that they regarded the risk as insignificant. That conclusion would in turn justify pervasive regulation limited only by the constraint of feasibility. In light of the fact that there are literally thousands of substances used in the workplace that have been identified as carcinogens or suspect carcinogens, the Government’s theory would give OSHA power to impose enormous costs that might produce little, if any, discernible benefit.

If the Government were correct in arguing that neither § 3(8) nor § 6(b)(5) requires that the risk from a toxic substance be quantified sufficiently to enable the Secretary to characterize it as significant in an understandable way, the statute would make such a “sweeping delegation of legislative power” that it might be unconstitutional under the Court’s reasoning in *Schechter Poultry* and *Panama Refining*. A construction of the statute that avoids this kind of open-ended grant should certainly be favored. . . .

. . . In this case the record makes it perfectly clear that the Secretary relied squarely on a special policy for carcinogens that imposed the burden on industry of proving the existence of a safe level of exposure, thereby avoiding the Secretary’s threshold responsibility of establishing the need for more stringent standards. In so interpreting his statutory authority, the Secretary exceeded his power. . . .

**MR. JUSTICE REHNQUIST, concurring in the judgment.**

. . . According to the Secretary . . . § 6(b)(5) imposes upon him an absolute duty, in regulating harmful substances like benzene for which no safe level is known, to set the standard for permissible exposure at the lowest level that “can be achieved at bearable cost with available technology.” While the Secretary does not attempt to refine the concept of “bearable cost,” he apparently believes that a proposed standard is economically feasible so long as its impact “will not be such as to threaten the financial welfare of the affected firms or the general economy.”

Respondents reply, and the lower court agreed, that § 6(b)(5) must be read in light of another provision in the same Act, § 3(8), . . . [and thus] requires the Secretary to demonstrate that any particular health standard is justifiable on the basis of a rough balancing of costs and benefits.
In considering these alternative interpretations, my colleagues manifest a good deal of uncertainty, and ultimately divide over whether the Secretary produced sufficient evidence that the proposed standard for benzene will result in any appreciable benefits at all. This uncertainty, I would suggest, is eminently justified, since I believe that this litigation presents the Court with what has to be one of the most difficult issues that could confront a decisionmaker: whether the statistical possibility of future deaths should ever be disregarded in light of the economic costs of preventing those deaths. I would also suggest that the widely varying positions advanced in the briefs of the parties and in the opinions [in this case] demonstrate, perhaps better than any other fact, that Congress, the governmental body best suited and most obligated to make the choice confronting us in this litigation, has improperly delegated that choice to the Secretary of Labor and, derivatively, to this Court.

I

In his Second Treatise of Civil Government, published in 1690, John Locke wrote that “[t]he power of the legislative, being derived from the people by a positive voluntary grant and institution, can be no other than what that positive grant conveyed, which being only to make laws, and not to make legislators, the legislative can have no power to transfer their authority of making laws and place it in other hands.” Two hundred years later, this Court expressly recognized the existence of and the necessity for limits on Congress’ ability to delegate its authority to representatives of the Executive Branch: “That Congress cannot delegate legislative power to the president is a principle universally recognized as vital to the integrity and maintenance of the system of government ordained by the Constitution.”

The rule against delegation of legislative power is not, however, so cardinal of principle as to allow for no exception. The Framers of the Constitution were practical statesmen, who saw that the doctrine of separation of powers was a two-sided coin. James Madison, in Federalist Paper No. 48, for example, recognized that while the division of authority among the various branches of government was a useful principle, “the degree of separation which the maxim requires, as essential to a free government, can never in practice be duly maintained.”

This Court also has recognized that a hermetic sealing-off of the three branches of government from one another could easily frustrate the establishment of a National Government capable of effectively exercising the substantive powers granted to the various branches by the Constitution. . . .

During the third and fourth decades of this century, this Court within a relatively short period of time struck down several Acts of Congress on the grounds that they exceeded the authority of Congress under the Commerce Clause or under the nondelegation principle of separation of powers, and at the same time struck down state statutes because they violated “substantive” due process or interfered with interstate commerce. When many of these decisions were later overruled, the principle that Congress could not simply transfer its legislative authority to the Executive fell under a cloud. Yet in my opinion decisions such as Panama Refining suffer from none of the excesses of judicial
policymaking that plagued some of the other decisions of that era. The many later decisions that have upheld congressional delegations of authority to the Executive Branch have done so largely on the theory that Congress may wish to exercise its authority in a particular field, but because the field is sufficiently technical, the ground to be covered sufficiently large, and the Members of Congress themselves not necessarily expert in the area in which they choose to legislate, the most that may be asked under the separation-of-powers doctrine is that Congress lay down the general policy and standards that animate the law, leaving the agency to refine those standards, “fill in the blanks,” or apply the standards to particular cases.

Viewing the legislation at issue here in light of these principles, I believe that it fails to pass muster. Read literally, the relevant portion of § 6(b)(5) is completely precatory, admonishing the Secretary to adopt the most protective standard if he can, but excusing him from that duty if he cannot. In the case of a hazardous substance for which a “safe” level is either unknown or impractical, the language of § 6(b)(5) gives the Secretary absolutely no indication where on the continuum of relative safety he should draw his line. Especially in light of the importance of the interests at stake, I have no doubt that the provision at issue, standing alone, would violate the doctrine against uncanalized delegations of legislative power. For me the remaining question, then, is whether additional standards are ascertainable from the legislative history or statutory context of § 6(b)(5) or, if not, whether such a standardless delegation was justifiable in light of the “inherent necessities” of the situation.

III

. . . [But] the legislative history contains nothing to indicate that the language “to the extent feasible” does anything other than render what had been a clear, if somewhat unrealistic, standard largely, if not entirely, precatory.1 . . .

. . . [And] there is little or nothing in the remaining provisions of the Occupational Safety and Health Act to provide specificity to the feasibility criterion in § 6(b)(5). It may be true, as suggested by Mr. JUSTICE MARSHALL, that the Act as a whole expresses a distinct preference for safety over dollars. But that expression of preference, as I read it, falls far short of the proposition that the Secretary must eliminate marginal or insignificant risks of material harm right down to an industry’s breaking point.

Nor [can the delegation be narrowed by reference to] a pre-existing administrative practice. Here, the Secretary’s approach to toxic substances like benzene could not have predated the enactment of § 6(b)(5) itself. Moreover, there are indications that the postenactment administrative practice has been less than uniform.

In some cases where broad delegations of power have been examined, this Court has upheld those delegations because of the delegatee’s residual authority over particular subjects of regulation. . . . In the present cases, however, neither the Executive Branch in

---

1 This sentence was moved from Part II.
general nor the Secretary in particular enjoys any independent authority over the subject matter at issue.

Finally, as indicated earlier, in some cases this Court has abided by a rule of necessity, upholding broad delegations of authority where it would be “unreasonable and impracticable to compel Congress to prescribe detailed rules” regarding a particular policy or situation. But no need for such an evasive standard as “feasibility” is apparent in the present cases. In drafting § 6(b)(5), Congress was faced with a clear, if difficult, choice between balancing statistical lives and industrial resources or authorizing the Secretary to elevate human life above all concerns save massive dislocation in an affected industry. . . . [But] Congress chose, intentionally or unintentionally, to pass this difficult choice on to the Secretary . . . .

IV

As formulated and enforced by this Court, the nondelegation doctrine serves three important functions. First, and most abstractly, it ensures to the extent consistent with orderly governmental administration that important choices of social policy are made by Congress, the branch of our Government most responsive to the popular will. Second, the doctrine guarantees that, to the extent Congress finds it necessary to delegate authority, it provides the recipient of that authority with an “intelligible principle” to guide the exercise of the delegated discretion. Third, and derivative of the second, the doctrine ensures that courts charged with reviewing the exercise of delegated legislative discretion will be able to test that exercise against ascertainable standards.

I believe the legislation at issue here fails on all three counts. The decision whether the law of diminishing returns should have any place in the regulation of toxic substances is quintessentially one of legislative policy. For Congress to pass that decision on to the Secretary in the manner it did violates, in my mind, John Locke’s caveat . . . that legislatures are to make laws, not legislators. Nor . . . do the provisions at issue or their legislative history provide the Secretary with any guidance that might lead him to his somewhat tentative conclusion that he must eliminate exposure to benzene as far as technologically and economically possible. Finally, I would suggest that the standard of “feasibility” renders meaningful judicial review impossible.

We ought not to shy away from our judicial duty to invalidate unconstitutional delegations of legislative authority solely out of concern that we should thereby reinvigorate discredited constitutional doctrines of the pre-New Deal era. If the nondelegation doctrine has fallen into the same desuetude as have substantive due process and restrictive interpretations of the Commerce Clause, it is, as one writer has phrased it, “a case of death by association.” . . .

If we are ever to resoulder the burden of ensuring that Congress itself make the critical policy decisions, these are surely the cases in which to do it. It is difficult to imagine a more obvious example of Congress simply avoiding a choice which was both fundamental for purposes of the statute and yet politically so divisive that the necessary decision or compromise was difficult, if not impossible, to hammer out in the legislative
forge. Far from detracting from the substantive authority of Congress, a declaration that the first sentence of § 6(b)(5) . . . constitutes an invalid delegation to the Secretary of Labor would preserve the authority of Congress. If Congress wishes to legislate in an area which it has not previously sought to enter, it will in today’s political world undoubtedly run into opposition no matter how the legislation is formulated. But that is the very essence of legislative authority under our system. It is the hard choices, and not the filling in of the blanks, which must be made by the elected representatives of the people. When fundamental policy decisions underlying important legislation about to be enacted are to be made, the buck stops with Congress and the President insofar as he exercises his constitutional role in the legislative process.

I would invalidate the first sentence of § 6(b)(5) . . . as it applies to any toxic substance or harmful physical agent for which a safe level, that is, a level at which “no employee will suffer material impairment of health or functional capacity even if such employee has regular exposure to [that hazard] for the period of his working life,” is, according to the Secretary, unknown or otherwise “infeasible.” Absent further congressional action, the Secretary would then have to choose, when acting pursuant to § 6(b)(5), between setting a safe standard or setting no standard at all. Accordingly, for the reasons stated above, I concur in the judgment of the Court . . . .

MR. JUSTICE MARSHALL, with whom MR. JUSTICE BRENNAN, MR. JUSTICE WHITE, and MR. JUSTICE BLACKMUN join, dissenting.

[Justice Marshall’s dissent argues that the government’s stricter reading of the statute is correct. The following text is taken from footnote 30:]

Finding obscurity in the word “feasible,” my Brother REHNQUIST invokes the nondelegation doctrine, which was last used to invalidate an Act of Congress in 1935. While my Brother REHNQUIST eloquently argues that there remains a place for such a doctrine in our jurisprudence, I am frankly puzzled as to why the issue is thought to be of any relevance here. The nondelegation doctrine is designed to assure that the most fundamental decisions will be made by Congress, the elected representatives of the people, rather than by administrators. Some minimal definiteness is therefore required in order for Congress to delegate its authority to administrative agencies.

Congress has been sufficiently definite here. The word “feasible” has a reasonably plain meaning, and its interpretation can be informed by other contexts in which Congress has used it. Since the term is placed in the same sentence with the “no employee will suffer” language, it is clear that “feasible” means technologically and economically achievable. Under the Act, the Secretary is afforded considerably more guidance than are other administrators acting under different regulatory statutes. In short, Congress has made “the critical policy decisions” in these cases.

The plurality’s apparent suggestion that the nondelegation doctrine might be violated if the Secretary were permitted to regulate definite but nonquantifiable risks is plainly wrong. Such a statute would be quite definite and would thus raise no constitutional question under Schechter Poultry. Moreover, Congress could rationally decide that it
would be better to require industry to bear “feasible” costs than to subject American workers to an indeterminate risk of cancer and other fatal diseases.


JUSTICE SCALIA delivered the opinion of the Court.

These cases present the . . . question[. . . ] whether § 109(b)(1) of the Clean Air Act (CAA) delegates legislative power to the Administrator of the Environmental Protection Agency (EPA). . . .

I

Section 109(a) of the CAA requires the Administrator of the EPA to promulgate [national ambient air quality standards (NAAQS)] for . . . air pollutant[s] . . . . Once a NAAQS has been promulgated, the Administrator must review the standard (and the criteria on which it is based) “at five-year intervals” and make “such revisions . . . as may be appropriate.” These cases arose when . . . the Administrator revised the NAAQS for particulate matter and ozone. American Trucking Associations . . . challenged the new standards in the [D.C.] Circuit . . . .

The [D.C.] Circuit . . . agreed . . . that § 109(b)(1) delegated legislative power to the Administrator in contravention of the . . . Constitution, Art. I, § 1, because it found that the EPA had interpreted the statute to provide no “intelligible principle” to guide the agency’s exercise of authority. The court thought, however, that the EPA could perhaps avoid the unconstitutional delegation by adopting a restrictive construction of § 109(b)(1), so instead of declaring the section unconstitutional the court remanded the NAAQS to the agency. . . .

III

Section 109(b)(1) of the CAA instructs the EPA to set “ambient air quality standards the attainment and maintenance of which in the judgment of the Administrator, . . . allowing an adequate margin of safety, are requisite to protect the public health.” The Court of Appeals held that this section as interpreted by the Administrator did not provide an “intelligible principle” to guide the EPA’s exercise of authority in setting NAAQS. “[The] EPA,” it said, “lack[ed] any determinate criteria for drawing lines. It has failed to state intelligibly how much is too much.” The court hence found that the EPA’s interpretation (but not the statute itself) violated the nondelegation doctrine. We disagree.

In a delegation challenge, the constitutional question is whether the statute has delegated legislative power to the agency. Article I, § 1, of the Constitution vests “[a]ll legislative Powers herein granted . . . in a Congress of the United States.” This text permits no delegation of those powers, and so we repeatedly have said that when Congress confers
decisionmaking authority upon agencies Congress must “lay down by legislative act an
intelligible principle to which the person or body authorized to [act] is directed to
conform.” We have never suggested that an agency can cure an unlawful delegation of
legislative power by adopting in its discretion a limiting construction of the statute. . . .
The idea that an agency can cure an unconstitutionally standardless delegation of power
by declining to exercise some of that power seems to us internally contradictory. The
very choice of which portion of the power to exercise—that is to say, the prescription of
the standard that Congress had omitted—would itself be an exercise of the forbidden
legislative authority. Whether the statute delegates legislative power is a question for the
courts, and an agency’s voluntary self-denial has no bearing upon the answer.

We agree with the Solicitor General that the text of § 109(b)(1) . . . at a minimum
requires that “. . . [the] EPA must establish uniform national standards at a level that is
requisite to protect public health from the adverse effects of the pollutant in the ambient
air.” Requisite, in turn, “mean[s] sufficient, but not more than necessary.” These limits on
the EPA’s discretion are strikingly similar to the ones we approved in Touby v. United
States (1991), which permitted the Attorney General to designate a drug as a controlled
substance for purposes of criminal drug enforcement if doing so was “‘necessary to avoid
an imminent hazard to the public safety.’” They also resemble the Occupational Safety
and Health Act of 1970 provision requiring the agency to “‘set the standard which most
adequately assures, to the extent feasible, on the basis of the best available evidence, that
no employee will suffer any impairment of health’”—which the Court upheld in [the
Benzene Case] (1980), and which even then-Justice REHNQUIST, who alone in that case
thought the statute violated the nondelegation doctrine, would have upheld if, like the
statute here, it did not permit economic costs to be considered.

The scope of discretion § 109(b)(1) allows is in fact well within the outer limits of our
nondelegation precedents. In the history of the Court we have found the requisite
“intelligible principle” lacking in only two statutes, one of which provided literally no
guidance for the exercise of discretion, and the other of which conferred authority to
regulate the entire economy on the basis of no more precise a standard than stimulating
the economy by assuring “fair competition.” See Panama Refining; Schechter Poultry.
We have, on the other hand, upheld [a statute] which gave the Securities and Exchange
Commission authority to modify the structure of holding company systems so as to
ensure that they are not “unduly or unnecessarily complicate[d]” and do not “unfairly or
inequitably distribute voting power among security holders.” We have approved the
wartime conferral of agency power to fix the prices of commodities at a level that “‘will
be generally fair and equitable and will effectuate the [in some respects conflicting]
purposes of th[e] Act.’” And we have found an “intelligible principle” in various statutes
authorizing regulation in the “public interest.” In short, we have “almost never felt
qualified to second-guess Congress regarding the permissible degree of policy judgment
that can be left to those executing or applying the law.”

It is true enough that the degree of agency discretion that is acceptable varies according
to the scope of the power congressionally conferred. While Congress need not provide
any direction to the EPA regarding the manner in which it is to define “country
elevators,” which are to be exempt from new-stationary-source regulations governing
grain elevators, it must provide substantial guidance on setting air standards that affect
the entire national economy. But even in sweeping regulatory schemes we have never
demanded, as the Court of Appeals did here, that statutes provide a “determinate
criterion” for saying “how much [of the regulated harm] is too much.” In *Touby*, for
example, we did not require the statute to decree how “imminent” was too imminent, or
how “necessary” was necessary enough, or even—most relevant here—how “hazardous”
was too hazardous. . . . It is therefore not conclusive for delegation purposes that . . .
ozone and particulate matter are “nonthreshold” pollutants that inflict a continuum of
adverse health effects at any airborne concentration greater than zero, and hence require
the EPA to make judgments of degree. “[A] certain degree of discretion, and thus of
lawmaking, inheres in most executive or judicial action.” Section 109(b)(1) . . . , which to
repeat we interpret as requiring the EPA to set air quality standards at the level that is
“requisite”—that is, not lower or higher than is necessary—to protect the public health
with an adequate margin of safety, fits comfortably within the scope of discretion
permitted by our precedent.

We therefore reverse the judgment of the Court of Appeals remanding for reinterpretation
that would avoid a supposed delegation of legislative power. . . .

**JUSTICE THOMAS, concurring.**

I agree with the majority . . . . I write separately, however, to express my concern that
there may nevertheless be a genuine constitutional problem with § 109, a problem which
the parties did not address.

The parties to these cases . . . wrangled over constitutional doctrine with barely a nod to
the text of the Constitution. Although this Court since 1928 has treated the “intelligible
principle” requirement as the only constitutional limit on congressional grants of power
to administrative agencies, see *J.W. Hampton, Jr., & Co.*, the Constitution does not speak
of “intelligible principles.” Rather, it speaks in much simpler terms: “All legislative
Powers herein granted shall be vested in a Congress.” U.S. Const., Art. 1, § 1 (emphasis
added). I am not convinced that the intelligible principle doctrine serves to prevent all
cessions of legislative power. I believe that there are cases in which the principle is
intelligible and yet the significance of the delegated decision is simply too great for the
decision to be called anything other than “legislative.”

As it is, none of the parties to these cases has examined the text of the Constitution or
asked us to reconsider our precedents on cessions of legislative power. On a future day,
however, I would be willing to address the question whether our delegation jurisprudence
has strayed too far from our Founders’ understanding of separation of powers.

**JUSTICE STEVENS, with whom JUSTICE SOUTER joins, concurring in part
and concurring in the judgment.**

. . . I wholeheartedly endorse the Court’s result and endorse its explanation of its reasons,
albeit with the following caveat.
The Court has two choices. We could choose to articulate our ultimate disposition of this issue by frankly acknowledging that the power delegated to the EPA is “legislative” but nevertheless conclude that the delegation is constitutional because adequately limited by the terms of the authorizing statute. Alternatively, we could pretend, as the Court does, that the authority delegated to the EPA is somehow not “legislative power.” Despite the fact that there is language in our opinions that supports the Court’s articulation of our holding, I am persuaded that it would be both wiser and more faithful to what we have actually done in delegation cases to admit that agency rulemaking authority is “legislative power.”

The proper characterization of governmental power should generally depend on the nature of the power, not on the identity of the person exercising it. If the NAAQS that the EPA promulgated had been prescribed by Congress, everyone would agree that those rules would be the product of an exercise of “legislative power.” The same characterization is appropriate when an agency exercises rulemaking authority pursuant to a permissible delegation from Congress.

My view is not only more faithful to normal English usage, but is also fully consistent with the text of the Constitution. In Article I, the Framers vested “All legislative Powers” in the Congress, just as in Article II they vested the “executive Power” in the President. Those provisions do not purport to limit the authority of either recipient of power to delegate authority to others. Surely the authority granted to members of the Cabinet and federal law enforcement agents is properly characterized as “Executive” even though not exercised by the President.

It seems clear that an executive agency’s exercise of rulemaking authority pursuant to a valid delegation from Congress is “legislative.” As long as the delegation provides a sufficiently intelligible principle, there is nothing inherently unconstitutional about it. Accordingly, . . . I would hold that when Congress enacted § 109, it effected a constitutional delegation of legislative power to the EPA.

§ 1.8. Notes on the later federal cases

1. The role of statutory interpretation. Recall note 2 in the Notes on the earlier federal cases, supra. Where Rehnquist would find that the statute delegates so broadly that it must be struck down, the majority chooses to read the statute in a more limited way so as to avoid an overly broad delegation. The majority’s approach is not universally respected. Judge Silberman of the D.C. Circuit, in a lower-court opinion in Am. Trucking, called that part of the Benzene opinion a “makeweight,” 195 F.3d 14, and accused the plurality of “analytically conflat[ing] the scope of the Secretary’s discretion . . . with the regulatory consequences of his interpretation of the statute.” The dissent, on the other hand, referring to past cases upholding broad delegations, finds no constitutional problem here, and therefore rejects not only Rehnquist’s blunt approach but also the majority’s more cautious approach.

Draft — do not distribute!
2. The historical pedigree of the non-delegation doctrine. Rehnquist traces the intellectual origins of the non-delegation doctrine back to Locke, and the Panama Refining Court traced the doctrine back to early 19th-century cases. Eric Posner and Adrian Vermeule dispute this account, arguing that there has never been a non-delegation doctrine—that the doctrine was invented in the late 19th century, was used without any originalist support in the 1930s, and should be abandoned entirely. Posner & Vermeule, Interring the Nondelegation Doctrine, 69 U. CHI. L. REV. 1721 (2002). But see Alexander & Prakash, Reports of the Nondelegation Doctrine’s Death Are Greatly Exaggerated, 70 U. CHI. L. REV. 1297 (2003), who argue that the doctrine really does have respectable historical roots.

3. The doctrine according to Rehnquist, Part I—congressional vs. agency responsiveness. Rehnquist gives three rationales for the non-delegation doctrine. The first is that policy decisions should be made by Congress, the branch with greater democratic accountability. Is this true? Agencies, after all, are run by the President, who is democratically accountable. Is there any theoretical reason to suppose that democratic accountability would be greater if we required more policy decisions to be made by Congress? Some argue that Congress uses delegation to give itself the credit for passing vague statutes while passing the blame for unpopular applications of the statute to the agency. See Schoenbrod, Power Without Responsibility: How Congress Abuses the People Through Delegation (1993). Posner & Vermeule respond that this proposition has neither theoretical nor empirical support; in particular, voters could punish Congress for vague delegations, anticipating the future unpopular applications. See also Mashaw, Prodelegation: Why Administrators Should Make Political Decisions, 1 J. L. Econ. & Org. 81 (1985).

4. The doctrine according to Rehnquist, Part II—the need for an intelligible principle. Rehnquist’s second rationale is that a delegate needs to be given an intelligible principle. Why is an intelligible principle necessary? One could argue that rule-of-law concerns (i.e., due process) demand it. Or one could argue (as the non-delegation doctrine traditionally does) that this is required for the delegate’s power to be truly executive and not legislative. Arguably, Rehnquist buys into the second conception. Does this make sense? Posner & Vermeule argue that any power wielded under a delegation by a statute is, by definition, executive. But see Alexander & Prakash, arguing that the non-delegation doctrine rests on a functional definition of “legislative” (making rules), not a formalistic one. (As an additional argument, consider Justice Stevens’s argument in Am. Trucking: that the Vesting Clause places legislative power in Congress doesn’t mean that Congress can’t delegate that power.)

5. The doctrine according to Rehnquist, Part III—facilitating judicial review. Rehnquist’s third rationale is that the non-delegation doctrine facilitates judicial review. Rehnquist calls this principle derivative of the second, but—as a free-floating concern with accountability—it can exist independently of the other two. For instance, in Am. Trucking, the lower court would have allowed the EPA itself to provide the intelligible principle, thus allowing legislative decisions to be made
by an administrative agency, but in a way that would promote accountability for the future. (The precedent for such a move can be found in Kenneth Culp Davis, *A New Approach to Delegation*, 36 U. Chi. L. Rev. 713 (1969), and *Amalgamated Meat Cutters & Butcher Workmen v. Connally* (D.D.C. 1971); see also *Int’l Union v. OSHA*, 938 F.2d 1310 (D.C. Cir. 1991).) Even in the Benzene case, Rehnquist looked to see whether a limiting principle could be found in agency practice.

Does this third rationale make sense as a justification for the non-delegation doctrine? Judicial review is not an end in itself, but a way of policing the standards that exist. If no standards exist—if the delegation really is so broad as to allow the President to act on a whim—judicial review serves no purpose. See Posner & Vermeule. The true concern is twofold: First, one might believe that some standards are necessary, in which case this third rationale folds into the second (recall that Rehnquist called it “derivative of the second”). Second, once some standards exist, one might want to make sure that the standards are being followed—but the concern that the President act within his authority sounds in procedural due process. See generally note 3 in the Notes on the earlier federal cases, *supra*.

6. **Legislative vs. executive essentialism.** Traditional accounts of the non-delegation doctrine explain that the executive must not be allowed to wield legislative power. This requires drawing a line between the two. Is this even possible? Any exercise of discretion is in a sense legislative. If one is pessimistic about the distinction between these two powers, there are various ways out: (1) Whenever the President acts under a delegation from Congress, he is exercising executive power, not legislative power; the non-delegation doctrine is a sham. This is Posner & Vermeule’s view. (2) Whenever the President is given rulemaking authority, he is exercising legislative power, but the “intelligible principle” test, rather than distinguishing between executive and legislative power, merely distinguishes between valid and invalid delegations of legislative power. This is Justice Stevens’s view, which retains the non-delegation doctrine in its current form, but changes its rhetoric.


**PHILLIPS, Chief Justice, delivered the opinion of the Court.**

... [T]he Texas Agriculture Code provides for the creation and operation of an “Official Cotton Growers’ Boll Weevil Eradication Foundation.” Subject to referendum approval from the affected cotton growers, this Foundation is authorized to operate boll weevil...
eradication programs and assess the growers for the cost. Appellees . . . , who are cotton growers subject to the Foundation’s jurisdiction, . . . challeng[ed] the Foundation’s assessments . . .

We . . . conclude . . . that the Legislature made an unconstitutionally broad delegation of authority to the Foundation, a private entity, thereby violating Article II, Section 1 of the Texas Constitution. . . .

I

A

. . . [T]he boll weevil[] presents a major economic threat to the Texas cotton industry. . . .

To aid in the ongoing battle against the boll weevil, the Legislature in 1993 authorized the creation of the Official Cotton Growers’ Boll Weevil Eradication Foundation. Instead of directly creating the Foundation, however, the Legislature merely authorized the Commissioner of Agriculture to certify some nonprofit organization representing cotton growers to create the Foundation and propose geographic eradication zones. The Act authorizes the creating organization or the Foundation to conduct referenda in each proposed eradication zone (“zone referenda”) to determine whether those cotton growers desire to establish an official boll weevil eradication zone. Contemporaneous with the zone referendum, the growers are also to elect a member to represent them on the Foundation’s board. If the growers vote not to establish a zone, their board selection is without effect.

Under the Act, once the initial zone has been created and the first board member elected, the growers of that zone must approve the assessment to fund the eradication at a subsequent referendum. Thereafter, the board is authorized to determine the assessment needed for each additional participating zone, which must be approved by the growers at a referendum. The Foundation may collect the assessment only if the assessment referendum passes. Approval of a zone and of the assessment each requires a vote of either two-thirds of the cotton growers in the zone or of those who farm more than one-half of the cotton acreage in the zone. The election of board members, on the other hand, requires only a plurality vote.

The Foundation exercises broad governmental powers. Besides being authorized to conduct elections in proposed eradication zones, the board may add an area to a zone under certain circumstances if approved by a referendum of cotton growers in the area. The board determines what eradication programs to conduct. The Foundation may impose penalties for late payment of assessments. A cotton grower who fails to pay an assessment within ten days of its due date must destroy his cotton crop. If the grower fails to do so, his crop is automatically declared a public nuisance. On the Foundation’s recommendation, and after notice, the Department of Agriculture must destroy it, even if not infested with boll weevils, at the owner’s cost. In addition, a cotton grower who violates the statute (including, presumably, by failing to pay an assessment or failing to destroy his own crop if payment is more than ten days late) is guilty of a Class C misdemeanor. Cotton which a delinquent grower has already produced and harvested is
subject to a lien. Representatives of the Foundation may enter private property which is subject to eradication without the owner’s permission for any purpose under the Act, including “the treatment, monitoring, and destruction of growing cotton or other host plants.” Finally, the Commissioner and the Foundation may adopt rules necessary to carry out the purposes of the Act.

While growers in a zone must approve their assessments, they do not approve the type of eradication program or the amount of debt incurred by the Foundation to finance it. These matters are left to the Foundation’s discretion. If the eradication program is discontinued for any reason, the Foundation may continue collecting assessments “as necessary to pay the financial obligations of the foundation.”

Under the Act, some power is retained by the Commissioner of Agriculture. For example, the Foundation can change the number of board positions or the eradication zone representation on the board only with the Commissioner’s approval. The Commissioner must also make rules to protect life and property from pesticides and other aspects of eradication programs. The Commissioner may prohibit planting cotton in zones when it would jeopardize the success of an eradication program. The Commissioner may exempt a cotton grower from payment of the Foundation’s assessment penalties if payment would leave the grower with less than $15,000 taxable income. The Foundation may expend revenue only on “programs approved by the commissioner as consistent with this subchapter and applicable provisions of the constitution.” Finally, the Commissioner must determine when elimination of boll weevils is no longer necessary to prevent economic loss to cotton growers.

After a referendum has passed, the cotton growers in the zone must be allowed to conduct referenda “periodically” under the terms prescribed in the initial referendum to determine whether to continue their assessments, although the Act says nothing about how often these referenda must occur. In addition, the Foundation must conduct a referendum on whether to discontinue the program on the petition of at least forty percent of the cotton growers in the zone. As noted, however, the Foundation may continue to collect assessments previously approved to pay its financial obligations. . . .

C

In May 1993, Texas Cotton Producers, Inc., a nonprofit organization representing cotton growers, petitioned the Commissioner for authority to create the Foundation. In its petition, Texas Cotton Producers proposed nine eradication zones around the state and provided that the Foundation board would consist of a corresponding nine members. Texas Cotton Producers further provided that “[o]n creation of the proposed official cotton growers’ boll weevil eradication foundation by TCP the initial board will be appointed by the board of TCP pending conduct of the board election.”

While the nine-member board is consistent with the requirements of [the statute], which permits a six, nine, twelve, or fifteen person board, the statute never authorizes the creating organization to appoint the initial Foundation board. Despite this flaw, the Commissioner certified Texas Cotton Producers to create the Foundation. In September
1993, Texas Cotton Producers incorporated the Foundation as a Texas nonprofit corporation, appointing nine members to the board who purportedly represented each of the nine proposed eradication districts. The Foundation has since conducted six zone referenda, and in each instance the growers elected as their board representative the person previously appointed by Texas Cotton Producers to represent that proposed zone. In the meantime, the remaining appointed members apparently voted on all Foundation matters, including the setting of assessments and expenditure of funds.

In April 1995, the Foundation conducted a referendum in the proposed High Plains Eradication Zone, which comprises all or parts of thirty West Texas counties. The cotton growers approved creation of the zone, and also approved [certain] maximum assessments. . . . [Another eradication zone was created, and then abolished, in the Lower Rio Grande Valley; growers in that area continued to be billed after the abolition of the program to retire the debt incurred.]

. . . [Some] cotton growers. . . challenged the validity of the referendum and assessments[, contending, among other things.] . . . that the Legislature improperly delegated authority to the Foundation in violation of the Texas Constitution’s separation of powers mandate. . . .

V

. . . [T]he growers contend that the Foundation is a private entity whose directors are neither constrained before they act by meaningful standards nor made accountable after they act by administrative, judicial, or popular review. In response, the Foundation contends that both the Legislature’s guidelines and the Commissioner of Agriculture’s supervisory authority are constitutionally adequate.

A

. . . The prohibition on unwarranted delegation of lawmaking power is “rooted in the principle of separation of powers that underlies our tripartite system of Government.” The United States Constitution expressly vests legislative power in the Congress, see U.S. Const. art. I, § 1, and the Texas Constitution similarly vests legislative power in our Legislature. Thus, “Congress is not permitted to abdicate or to transfer to others the essential legislative functions with which it is vested.” Schechter Poultry. Likewise, in our State “[t]he power to pass laws rests with the Legislature, and that power cannot be delegated to some commission or other tribunal.”

Yet, like many truisms, these blanket pronouncements should not be read too literally. Even in a simple society, a legislative body would be hard put to contend with every detail involved in carrying out its laws; in a complex society it is absolutely impossible to do so. Hence, legislative delegation of power to enforce and apply law is both necessary and proper. Such power must almost always be exercised with a certain amount of discretion, and at times the line between making laws and enforcing them may blur. As Justice Scalia has observed:
Once it is conceded, as it must be, that no statute can be entirely precise, and that some judgments, even some judgments involving policy considerations, must be left to the officers executing the law and to the judges applying it, the debate over unconstitutional delegation becomes a debate not over a point of principle but over a question of degree. . . . [T]he limits of delegation “must be fixed according to common sense and the inherent necessities of the governmental co-ordination.” Since Congress is no less endowed with common sense than we are, and better equipped to inform itself of the “necessities” of government; and since the factors bearing upon those necessities are both multifarious and (in the nonpartisan sense) highly political . . . it is small wonder that we have almost never felt qualified to second-guess Congress regarding the permissible degree of policy judgments that can be left to those executing or applying the law.

[Though one should avoid] “allowing delegation of power to exercise unguided discretion in individual cases,” . . . “the kind of government we have developed could not operate without” allowing legislatures to delegate rulemaking authority to administrative bodies. [A]t the height of the New Deal, a young Louis Jaffe asserted:

It was said by the courts for many years that Congress could not “delegate” its powers to administrative officers, though it could give them power “to fill up the details.” But that use of language has worn thin and become rather preposterous, and it is now admitted that Congress may “delegate” power to public officers, since there is no specific constitutional prohibition against delegation as such, provided the delegation is properly prescribed and is felt to be necessary to the operation of government. In the same way we may restate the question in this field as one of “proper delegation” or “reasonable delegation,” or we might drop the word “delegation” completely, at least as indicating a constitutional category, and regard the question simply as one of reasonableness within the due process clause.\(^\text{10}\)

Even before the Depression, one state court noted: “It only leads to confusion and error to say that the power to fill up the details and promulgate rules and regulations is not legislative power.”

Even in its heyday, the nondelegation doctrine was sparingly applied, having been used by the United States Supreme Court to strike down a federal statute only three times. See *Panama Refining; Schechter Poultry; Carter Coal*. Since the Court retreated from its opposition to New Deal initiatives, it has consistently upheld congressional delegations.

---

\(^{10}\) Delegations have sometimes been attacked on substantive due process or due course grounds, rather than under separation of powers, on the theory that the statute or ordinance allows the delegate, whether it be an administrative agency or a segment of the population, to exercise governmental power arbitrarily. . . .

More commonly, however, Texas has rooted its delegation jurisprudence only in the principle of separation of powers. Because we find that the delegation of authority to the Foundation violates the principle of separation of powers, we do not reach the question of whether it may also violate federal substantive due process.
Texas courts have also generally upheld legislative delegations to state or municipal agencies. . . .

But there are some indications that extreme judicial deference to legislative delegation may be declining. A number of Supreme Court justices have emphasized the need for adequate legislative standards. For example, Justice Brennan has cautioned: “Formulation of policy is a legislature’s primary responsibility, entrusted to it by the electorate, and to the extent Congress delegates authority under indefinite standards, this policy-making function is passed on to other agencies, often not answerable or responsive in the same degree to the people.” Moreover, “[m]any distinguished scholars and judges [have become] so concerned about the enormous discretionary power of agencies that they [have] urged reinvigoration of the doctrine.”

. . . [T]his Court has been especially willing to strike down delegations of legislative authority to the judicial department. And [delegations to administrative agencies have also been viewed skeptically in various Texas cases.]

B

As difficult as the issue of proper legislative delegation may be, the considerations are even more complex when the delegation is made not to another department or agency of government, but to a private individual or group. While at first blush such delegations might seem manifestly unconstitutional, further reflection demonstrates that they also are frequently necessary and desirable. Presumably no one would argue that the state should not accord the full benefits and responsibilities of a marital union to a couple who was married by a minister, priest, or rabbi rather than a judge. Also, the delegation of authority to private associations to promulgate certain industrial and professional standards has been of immense benefit to the public. For example, a number of states have adopted existing or future versions of the National Electrical Code, promulgated by an industry association, “turning a technical and complex task often quite beyond the competence of many city councils or even state legislatures over to a specialized private group.”

Still, private delegations clearly raise even more troubling constitutional issues than their public counterparts. On a practical basis, the private delegate may have a personal or pecuniary interest which is inconsistent with or repugnant to the public interest to be served. More fundamentally, the basic concept of democratic rule under a republican form of government is compromised when public powers are abandoned to those who are neither elected by the people, appointed by a public official or entity, nor employed by the government. Thus, we believe it axiomatic that courts should subject private delegations to a more searching scrutiny than their public counterparts.

While the United States Supreme Court has upheld many statutes involving some degree of private delegation, see, e.g., *Sunshine Anthracite Coal; Currin* [and others], state courts have frequently invalidated such provisions. Unfortunately, scholars have concluded that these cases do not yet, when taken together, evince a coherent constitutional standard. When Professor Davis issued the second edition of his treatise,
for example, he abandoned his earlier effort to analyze the state law on private
delegations “because identifiable principles do not emerge.” We thus begin our analysis
with full recognition that, if the delegation at issue is to a private entity, we must craft our
own criteria to judge its constitutionality.

C

We first address whether the Foundation is a public or private entity for purposes of the
nondelegation doctrine. Before ever being subject to its authority, any grower will
already have had the right to participate in one or more referenda deciding whether to
ratify the zone, who to elect to the board, and what amount is to be assessed on the
grower’s acreage. These referenda are not purely private affairs, being conducted
according to state law, under rather extensive regulations promulgated by the
Commissioner of Agriculture. For example, the Commissioner approves the ballots,
verifies the results, and issues certificates of election to the prevailing board candidates.
Yet they are not popular elections either, as the suffrage is strictly limited to eligible
growers.

Similarly, the statutory provisions as to governmental powers suggest both public and
private attributes. The Act exempts the Foundation from taxation and affords state
indemnification to its board members. The Foundation’s board members, officers, and
employees have official immunity except for gross negligence, criminal conduct, or
dishonesty. The Foundation must adopt and publish its rules in accordance with state
requirements, it may be dissolved by the Commissioner when its purpose has been
fulfilled, and it (or at least its board) is subject to . . . the Texas Sunset Act. The
Legislature specifically denominates the Foundation a “governmental unit” for purposes
of immunity from suit under the Tort Claims Act. Finally, the Foundation does not
dispute that it is a “governmental body” subject to the Texas Open Meetings Act.

For many purposes, however, the Foundation is not a state agency. Thus, the funds the
Foundation collects are expressly “not state funds and are not required to be deposited in
the state treasury.” The Act also does not subject the Foundation to state purchasing or
audit requirements, and its board members are not required to take oaths of office.
Finally, there is no provision for administrative appeal from Foundation decisions, except
as to penalties imposed for nonpayment of assessments.

In sum, we do not find it easy to categorize the Foundation as either a public or private
agency, a difficulty that may exist with many contemporary bodies. However, courts
have universally treated a delegation as private where “interested groups have been given
authoritative powers of determination, usually in conjunction with a public administrative
agency.” That the decisionmakers are elected, that their decisions affect only those they
represent, that they exercise police power, or that the government constrains their power
by advance restriction or subsequent review may all, as we shall see, be relevant in
assessing the validity of the private delegation; but they do not keep it from being
considered private in nature. Because the Act delegates authoritative power to private
interested parties, we conclude that it is a private entity for purposes of applying the
nondelegation doctrine.
Now we must determine what standard to apply in determining whether the private delegation was appropriate. Because of the additional risks posed by such delegations to the proper separation of governmental powers, a number of factors should be considered by a reviewing court. . . . [After canvassing various commentators’ suggested lists of factors to consider, the Court settles on the following eight:]

1. Are the private delegate’s actions subject to meaningful review by a state agency or other branch of state government?

2. Are the persons affected by the private delegate’s actions adequately represented in the decisionmaking process?

3. Is the private delegate’s power limited to making rules, or does the delegate also apply the law to particular individuals?

4. Does the private delegate have a pecuniary or other personal interest that may conflict with his or her public function?

5. Is the private delegate empowered to define criminal acts or impose criminal sanctions?

6. Is the delegation narrow in duration, extent, and subject matter?

7. Does the private delegate possess special qualifications or training for the task delegated to it?

8. Has the Legislature provided sufficient standards to guide the private delegate in its work?

We emphasize at the outset that these standards apply only to private delegations, not to the usual delegation by the Legislature to an agency or another department of government. . . . Furthermore, nothing in our analysis should be read as suggesting what provisions of the Act would or would not pass muster were this a public delegation. Likewise, we express no opinion as to whether any of the other statutory enactments cited by the dissenting justices would or would not pass constitutional muster. Thus, in no way does our opinion, as the dissenting justices fear, “ultimately threaten the heretofore established role of quasi-governmental entities under Texas law” . . . .

First, while the Foundation is subject to some oversight by the Commissioner of Agriculture, the review is uneven and incomplete. . . . The Commissioner is required to adopt rules for the zone referenda and board elections, rules specifying hardship exemptions from assessment penalties, and rules “to protect individuals, livestock, wildlife, and honeybee colonies” in eradication areas. . . . [T]he Commissioner’s regulations relating to the protection of human life and the environment are fairly extensive. The Act also provides that the Commissioner may adopt rules regulating
cotton planting in eradication zones and adopting a schedule of penalty fees, which he has also done.

The Commissioner could not, however, adopt any procedure for reviewing such critical decisions as the amount of assessments adopted by the growers, the total amount of funds expended on eradication, the amount of debt incurred by the Foundation, or the repayment terms for such debts. . . . The dissenting justices rely on the proviso that the Foundation must expend revenue only on “programs approved by the commissioner as consistent with this subchapter,” but “programs” cannot be stretched to include these critical factors. These determinations are left exclusively to the growers and the board, excluding the Commissioner or any state agency from meaningful review.

Finally, contrary to the dissenting justices’ conclusion, the Commissioner has no general authority to “revoke the Foundation’s certification” if it fails to comply with the procedural provisions of the Act. The Act provides that “[t]he commissioner shall certify the petitioning organization selected under . . . this code as the organization authorized to create an official boll weevil eradication foundation . . . . The commissioner may revoke the organization’s certification on 60 days written notice if the organization fails to meet the requirements of this subchapter.” It is clear that this section authorizes the Commissioner to revoke only the authority of the creating organization. Once the Foundation is created as an independent entity, the Commissioner has no authority to dissolve it, except when its eradication purpose has been fulfilled or it has become inoperative and abandoned. Thus, the first factor weighs against the delegation.

Judging the statute as it is written, rather than as it operated in practice, the second factor militates in favor of the private delegation. The growers in each zone are allowed to vote on whether to participate in the eradication program, and thereby subject themselves to the Foundation’s jurisdiction, and are allowed to approve or reject any proposed assessment. Although the Foundation in actuality operated for nineteen months with a board controlled by Texas Cotton Producers’ appointees, this process was inconsistent with the statutory contemplation that the Foundation should at all times by governed by the elected board. We thus do not consider that actual operation in reviewing the constitutionality of the Act. 11

The third factor weighs against the delegation. Far from merely devising eradication guidelines, the Foundation actually applied the programs it devised to all growers in zones where the program was approved. In accordance with its statutory authority, the Foundation collected assessments from individual growers and entered those growers’ property to carry out its eradication programs.

The fourth factor also weighs against the delegation. The Foundation board members are cotton growers who have a direct pecuniary interest in the eradication programs implemented by the Foundation.

11 The dissenting justices’ conclusion that we have based our decision in part on the Commissioner’s approval of an appointed board, post at n.4, is thus erroneous.
Under the circumstances of this case, the fifth factor does not weigh in our consideration of whether the statute as a whole is an unconstitutional delegation of authority to the Foundation. The Foundation is vested with authority to impose monetary penalties for late payment of the assessments and to direct the Department to destroy a delinquent growers’ crops, and it is further empowered to adopt rules, a violation of which is a criminal offense. While this authority to impose penal sanctions strongly suggests an improper private delegation, principles of severability would allow us to strike down this power and still uphold the Act. . . . [T]he assessment and expenditure provisions of the Act could be implemented without the penalty provisions. Thus, even though the penalty provisions seem to represent an unconstitutional delegation of authority to the Foundation, this should not weigh in judging the validity of the Foundation’s core function under the Act, i.e., the levying and collecting of assessments and the expenditure of those assessments on eradication programs.

The sixth factor is inconclusive under the circumstances of this case. While the statute pertains to a specific, narrow purpose—eradication of the boll weevil and other cotton pests—it does not limit the program’s cost and duration, other than to provide that the program is subject to the Sunset Act and that it should be discontinued once the boll weevil is eradicated.

The seventh factor, on its face, weighs against the delegation. While the Act is designed to allow those with firsthand experience in the cotton industry to lead the eradication effort, there is no assurance that those elected will actually have special qualifications or training regarding eradication of boll weevils . . . . The facts are thus quite distinct from a private delegation, say, for the promulgation of a municipal electrical code by an industry association consisting of electrical contractors, inspectors, and manufacturers. There is, of course, some tension between this factor and the second factor. It would ordinarily be difficult for a private delegation both to guarantee adequate representation of those affected by the delegation and to vest decisionmaking authority in a group of experts. There is no evidence in the record of a disinterested and yet eminent pre-existing entity to which the devising and implementing of a boll weevil eradication program could have been delegated. Thus, while the Act fails to meet the seventh factor, this failure is excused by the satisfaction of the second factor.

Finally, the eighth factor weighs against the delegation. The Legislature has provided very few statutory standards to guide the Foundation. While the Act provides the procedures for zone referenda and specifies the powers and duties of the board, it provides no guidance as to how assessments are to be set or the amount of debt that the Foundation may incur. Thus, in practice, the Foundation had free rein to incur over $9 million in debt in the Lower Rio Grande Valley Zone to be repaid by the growers there through several years of assessments, even though those growers voted within 21 months to discontinue their eradication program.

We recognize that the judicial branch should defer to the judgment of the people’s elected representatives whenever possible, and we by no means suggest that a private delegation must satisfy all eight of these factors. We recognize also that courts should, when possible, read delegations narrowly to uphold their validity. . . . Here, however, the
invalidity of the delegation does not hinge on any one provision of the Act that might be narrowly interpreted; rather, the Act as a whole represents an overly broad delegation of legislative authority to a private entity, violating a majority of the eight factors we have set forth.12 Therefore, the Act cannot stand. . . .

HECHT, Justice, concurring in part and dissenting in part, and concurring in the judgment.

The Texas Boll Weevil Eradication Foundation wields more legislative power with less restraint than any other privately chartered nonprofit corporation in Texas, or, as far as I can tell, in the history of Texas. Twelve other states have private foundations to help eradicate boll weevils; none has—or appears to need—the power the Legislature has ceded the Texas Boll Weevil Eradication Foundation. The Foundation is by no means a typical administrative agency; it is not even an atypical administrative agency; it is a complete anomaly in the structure of government.

Thus, I agree with the Court that the Official Cotton Growers’ Boll Weevil Eradication Foundation Act, delegates legislative power to the Foundation in violation of Article II, Section 1 of the Texas Constitution, but I think this conclusion is much clearer than the Court does. The Court’s decision certainly does not “threaten the heretofore established role of quasi-governmental entities under Texas law,” as the sky-is-falling rhetoric of Justice Cornyn’s opinion forebodes. Holding the extraordinary delegation of power to the Foundation unconstitutional is no more a threat to other administrative agencies than it is to the ozone layer.

The vice in the delegation to the Foundation does not lie simply in the fact that the Foundation is a private entity or that, as I conclude, it is privately managed. Nor is the vice simply that the only standards prescribed by the Legislature to guide the Foundation are extremely broad, or that the Foundation is largely free of public supervision. . . . Plainly put, the Texas Boll Weevil Eradication Foundation is little more than a posse . . . . The Foundation is only a private, nonprofit corporation, run mostly by individuals who have never been elected or appointed by an elected official, but the Legislature has empowered it to conduct elections of cotton growers in various areas throughout the State, to determine whether and how to conduct a boll weevil eradication program, to assess cotton farmers millions of dollars to pay for the program, to deposit those assessments in its own account not subject to state purchasing and auditing requirements, to enter private property and destroy healthy crops for failure to pay assessments, to borrow unlimited amounts and commit cotton farmers to repay its debts without their approval, and to spend the money without any state supervision. The Legislature has broad discretion to delegate authority to administrators, but for reasons the Court explains, this goes too far. . . .

---

12 We thus disagree with the dissent’s conclusion that we have not accorded the Act a presumption of constitutionality.
The Court’s reading of the Act makes the delegation issue closer than it should be. If one reads the Act as everyone involved in implementing it has[,] . . . the unconstitutional delegation of power is clear. As far as I can determine, the Legislature has never made a similar delegation of authority, and it was not necessary to do so in this situation.

A

. . . [T]he Act provides:

(a) A nonprofit organization authorized under the laws of this state that represents cotton growers may petition the commissioner [of agriculture] for certification as the organization authorized to:

(1) create a foundation;

(2) conduct the initial election of the board; and

(3) conduct referenda to establish eradication zones.

(b) A petition under this section must include:

(1) a geographic description of each proposed eradication zone, including a separate proposed eradication zone for the High Plains Boll Weevil Suppression Program Area and the St. Lawrence Cotton Growers Boll Weevil Control Zone;

(2) an initial plan for representation for each proposed eradication zone on a board consisting of 6, 9, 12, or 15 members; and

(3) any other information required by the commissioner.

(c) Not later than the 60th day after the date on which the commissioner receives a petition for certification, the commissioner shall hold a public hearing to consider the pending petition.

(d) After a hearing is held under Subsection (c) of this section the commissioner may select one organization to implement this subchapter and shall certify that the selected organization:

(1) has submitted a petition that complies with this subchapter;

(2) can adequately represent the interests of cotton growers in the proposed eradication zones described by the organization’s petition; and

(3) is authorized to conduct eradication zone referenda and initial board elections under Sections 74.105 and 74.106 of this code.
The Act requires the Commissioner to certify the selected organization “as the organization authorized to create an official boll weevil eradication foundation.”

The Act does not specify what type of entity the foundation is to be[, but everyone involved seems to have recognized that the foundation was a nonprofit corporation]. . . .

Unless a nonprofit corporation is to be governed by its members (if it has members), it must be governed by a board of directors, and the initial directors must be named in its articles of incorporation. The Court reads the Act to contemplate that the foundation must “at all times” be governed by a board of directors, each of whom has been elected. But the Act nowhere states that all directors of the foundation must be elected. To the contrary, the Act strongly suggests that the foundation may be governed by non-elected directors. . . . [A]n organization petitioning the Commissioner for certification to create the foundation must present “an initial plan for representation for each proposed eradication zone on a board consisting of 6, 9, 12, or 15 members” (emphasis added). Proposed zones cannot have elected board members because a board election must “be held concurrently with an eradication zone referendum,” and “[i]f a referendum to establish an eradication zone fails, the concurrent election of board members from the proposed eradication zone . . . has no effect.” Thus, only established zones can have elected board members. A plan for representing proposed zones on the Foundation board must necessarily call for non-elected directors.

It is not even clear that every established zone must be represented by an elected board member. The Act says only that “[e]ach zone shall be represented on the board.” The Act requires the zone referendum ballot to state “whether a board member is elected by a plurality or a majority of the votes cast.” (The Commissioner has determined that board members are elected by a plurality of the votes cast.) If a zone is established by referendum but no one is elected by a plurality or majority, the Act is silent on how the zone’s board representative is to be chosen. (The Commissioner has not promulgated a rule to cover such a situation, and the Foundation has promulgated no rules at all.)

The Court’s conclusion that the Foundation must “at all times” be governed by elected board members appears to be based on several references to “initial elections” . . . . The Court appears to infer from references to initial elections that the initial directors must all be elected, but the inference is not compelled by logic and is contrary to the express language in the statute that the petitioning organization present a plan for representation of proposed eradication zones.

Moreover, the Court’s reading of the Act is contrary to the Commissioner’s rules. . . . The Court’s reading of the Act is also contrary to that of every other person or entity involved with its implementation. . . .

The Foundation operated for more than five months before the first zone was established by referendum and the first board member elected. During this time the board members appointed by Texas Cotton Producers met to conduct the Foundation’s business and plan for the first referendum. Legal counsel from the Department of Agriculture met with the board to advise them concerning its legal responsibilities. Counsel never raised a concern
about the non-elected members on the board, thus indicating counsel’s approval of the
constitution of the board. . . .

When the 74th Legislature convened in January 1995, a majority of the Foundation’s
board were still members appointed by Texas Cotton Growers and not elected. When the
session ended, four members of the board still had not been elected. . . .

It is possible, of course, that the Court is correct and the Commissioner, Department, and
Foundation, and at least apparently the Legislature, are all wrong about whether non-
elected members can serve on the Foundation board. But when the Court’s construction
of the Act also contradicts strong suggestions in the Act itself and all operations under the
Act since its inception, the Court’s position is virtually untenable. While it is possible to
read the Act to require only elected board members, that construction is less reasonable
or plausible than the opposite one.

B

I join in the Court’s opinion concerning delegation because even under its construction of
the Act, I agree that the delegation of power to the Foundation is unconstitutional. If the
Act allows the Foundation to operate with a non-elected board, the delegation of power to
the Foundation is even less defensible than the Court believes. The Court details the
respective powers of the Foundation, the Commissioner, and the Department. Without
any supervision by an elected official or governmental agency, the Foundation can wield
enormous power. For example, while the Foundation must “conduct a referendum in each
proposed eradication zone to determine whether cotton growers desire to establish an
official boll weevil or pink bollworm eradication zone,” the statute imposes no deadlines
on this duty. In the more than four years the Foundation has been in existence, it has
conducted zone referenda in only six of the proposed zones. The Foundation thus has the
unilateral right to determine whether, over a course of years, at least, to poll cotton
growers to determine whether a zone should exist. The Foundation also has the unilateral
right to determine the maximum assessment to be put to a referendum. The voters cannot
raise or lower this amount; they can only vote for or against it.

While assessments must “be used solely to finance programs approved by the
commissioner as consistent with” the Act and the Constitution, the Foundation has great
leeway in determining the details of those programs. The nature of the programs and the
manner in which they are conducted can profoundly affect cotton growers, as the
experience in the Lower Rio Grande Valley zone, which I describe below, indicates.
Moreover, the Foundation has the absolute right to determine how much of the cost of a
program to borrow and how much to pay for with collected revenue. The Foundation can
borrow and spend the maximum total assessment in the first few months of the program
and obligate cotton growers to repay the debt over the total period approved in the
referendum. If the program appears to be unsuccessful, growers have no effective means
of stopping it before the total assessment has been committed.

The Foundation’s powers to enforce collection of assessments are truly draconian. The
Foundation may enter private property essentially at will and without the owner’s
permission at any time during daylight hours. It may set penalties for late payments. The Court observes that growers are entitled to protest imposition of such penalties to the Department, but as a very technical matter, the Court is in error. [A] Commissioner’s rule . . . states that “[a] person against whom the department has assessed a penalty . . . may protest such action with the commissioner.” But the Act expressly authorizes the board, not the Department, to set penalties. Furthermore, the Act requires the Department—it has no discretion—to destroy crops on the Foundation’s recommendation, even if they are not infested with boll weevils. The assertion in Justice Cornyn’s opinion that “the statute itself . . . and not the Board . . . determines the circumstances under which a grower’s crop may be destroyed for nonpayment of assessments” is, I believe, very misleading. The Act “determines” only one circumstance: “on recommendation of the foundation.” The only limit the Act places on the Foundation’s power to destroy crops is that it cannot do so if it does not feel it should. Otherwise, the Foundation has complete discretion. . . .

The potential power delegated to the Foundation and its actual exercise both demonstrate the unconstitutionality of the Act. . . . The Foundation and its agents are immune from liability for all their acts except gross negligence, criminal conduct, and dishonesty, and the Foundation has sovereign immunity. Applying the Court’s test to this construction of the Act, as opposed to the Court’s much more deferential reading, the result is even clearer.

C

No other private agency in Texas possesses such unrestricted power. The only entity remotely similar is the Texas Automobile Insurance Plan Association. TAIPA is a nonprofit corporation whose members are all authorized automobile insurers. Its 15-member governing body consists of eight members elected by the member insurers, five public members nominated by the Office of Public Insurance Counsel and selected by the Commissioner of Insurance, and two local recording agents. TAIPA, like the Texas Boll Weevil Eradication Foundation, is thus a private entity with a largely private governing body. Unlike the Foundation, however, TAIPA’s authority is very limited. It can only propose a plan for assigned risk insurance to the Commissioner for his approval, and administer it. The statute imposes certain requirements on the plan. The Commissioner retains the power to set rates under the plan. Given the detailed statutory requirements and the supervisory role of the Commissioner, TAIPA is little more than an adviser to and implementer for the Commissioner.

Justice Cornyn’s opinion suggests that the delegation of power to the Texas Boll Weevil Eradication Foundation is no different from statutes that condition licensing on graduation from accredited schools. The argument is that such statutes give accrediting associations power over licensing. But any power accrediting associations have is directed at the schools, not at licensing in Texas, and derives from association members, not from the Texas Legislature. The limited operation of such associations simply cannot be compared with the Foundation’s broad powers.
I am not aware of any entity in Texas’ history remotely resembling the Texas Boll Weevil Eradication Foundation. The need for an unprecedented delegation of power to the Foundation is not apparent. Twelve other states have boll weevil eradication laws. . . . No other state has determined that a viable boll weevil eradication program depends upon so expansive a delegation of power to a private entity as Texas gives its Boll Weevil Foundation. The Foundation is thus uniquely situated, both in Texas government, and among similar organizations in other states.

D

Because I conclude that the Texas Boll Weevil Eradication Foundation is as the Legislature intended, a private nonprofit corporation privately managed, and that the Foundation has been delegated broad, unrestricted legislative power, I agree with the Court that the Act violates Article II, Section 1 of the Texas Constitution. I also agree with the Court’s explanation why the use of referenda does not save the Act, but I offer a few additional reasons.

[Justice Hecht summarizes the holding in Carter Coal.] The Court has used a similar analysis to strike down other legislation. In Eubank v. City of Richmond (1912), the Court invalidated a city ordinance that allowed the owners of two-thirds of the property abutting a street to establish building setback lines. The evil the Court attacked was the delegated power of some property owners “to virtually control and dispose of the proper rights of others” without any “standard by which the power thus given is to be exercised.” In Washington v. Roberge (1928), the Court struck down an ordinance which required the approval of owners of two-thirds of the property within 400 feet of a proposed home for the aged poor before the facility could be constructed. Again, the Court objected that the ordinance was “uncontrolled by any standard or rule prescribed by legislative action.” Property owners, the Court observed, were “not bound by any official duty, but [were] free to withhold consent for selfish reasons or arbitrarily.”

The Court cited both these cases with approval in City of Eastlake v. Forest City Enterprises, Inc. (1976). There the Court upheld a city charter provision requiring proposed land use changes to be ratified by 55% of the people voting at a city-wide referendum. A referendum to a legislative body’s entire constituency cannot, the Court said, “be characterized as a delegation of power.” “In establishing legislative bodies, the people can reserve to themselves power to deal directly with matters which might otherwise be assigned to the legislature.” This was far different from “the standardless delegation of power to a limited group of property owners condemned by the Court in Eubank and Roberge.”

Eubank and Roberge do not preclude a legislative body from conditioning regulation on local approval. [See Currin.] [A] referendum does not involve an improper delegation of governmental power if it is to the entire constituency or if it is simply a condition of the application of legislatively determined regulation. A referendum does involve a delegation of power when the voters of a group themselves determine what the law or regulation is to be, as well as whether it is to be applied to them. Such a delegation has never been approved by the Supreme Court.
The flaw with referenda in the Boll Weevil Act is similar to that identified in *Carter*: the Act places decisions affecting cotton growers not in the hands of the public or its government, but in the hands of a few cotton growers. A referendum under the Act carries when approved by two-thirds of those voting or by those farming more than half the cotton acreage in the zone. In all but one instance one-third or less of the cotton growers in a zone imposed assessments on all the other cotton growers in the zone. *Carter* condemned as “legislative delegation in its most obnoxious form” the power of a majority in one industry to control the affairs of an unwilling minority. Under the Act before us, a minority of cotton growers can control the affairs of an unwilling majority.

The problem is not simply that referenda may carry on the vote of a very small group of cotton growers. The outcomes of many elections are determined by a minority of voters. If the issue were whom to choose as a representative of the people, or whether to accept regulation proposed by the people’s representatives, as in *Currin*, the fact that a small number of people voted in the election would not be of constitutional significance. But when the electorate consists not of all the citizens in an area but of the members of one industry, the referendum may be very unfair.

For example, the Act contemplates that when a zone is established and an assessment approved, every cotton grower in the zone must pay the assessment, whether his cotton is infested with boll weevils or not. The reason for this scheme appears to be that boll weevils cannot be eradicated except over a large area; one farmer may eliminate the insects from his field, but they may simply move to his neighbor’s field. Once the first farmer ceases his efforts, the insects may return. However, farmers who are barely surviving economically must weigh the risk of infestation in deciding whether to spend limited resources for pesticides and other efforts. The Act allows those in a stronger financial position to force those in a weaker position to contribute to eradication programs when they otherwise would not do so; conversely, it also allows growers in a weaker position to prevent all concerted eradication efforts.

Thus, I agree with the Court that the delegation of power to the Foundation is unconstitutional. The use of referenda does not save it.

**CORNYN, Justice, joined by ENOCH, SPECTOR and ABBOTT, Justices, . . . dissenting . . . .**

. . . In striking down this statute, the Court commits four major errors that ultimately threaten the heretofore established role of quasi-governmental entities under Texas law. Repeatedly, we have recognized that legislative delegations of authority are an essential part of modern governance in an increasingly complex society. The Court’s decision in this case, ironically enough, tends to prove rather than refute the arguments of those administrative law scholars who contend that state delegation cases lack a defining principle.

* In relevant part.
That is not to say that the concerns embodied in the nondelegation doctrine are trivial; just the opposite is true. There can be no doubt that when lawmaking power is delegated to an administrative agency to carry out the Legislature’s policy goals, that agency is necessarily given the power to give substance to that legislative policy. But . . . given proper safeguards, this delegation is both proper and necessary . . . .

My fundamental objection to the Court’s approach is that while the nondelegation doctrine has undergone at least five stages of development in our national jurisprudence [ (“the Early Stage, the Public Interest Stage, the ‘Strict’ Standards Stage, the ‘Loose’ Standards Stage, and the Procedural Safeguards Stage”)], the Court’s approach instead represents a stage of arrested development. The Court’s extraordinary skepticism of this particular delegation emanates from two Supreme Court cases, dating back to the Court’s “Strict” Standards Stage during the first half of this century, both of which, although never expressly overruled, have been limited to their facts. Not surprisingly, this strict nondelegation approach is followed by only a minority of the states. It is on this flimsy foundation that the Court’s public policy arguments are erected. . . .

. . . [T]he Court [also] strains to conclude that the Texas Boll Weevil Eradication Foundation is a “private” entity. Characterizing the Foundation as a “private” agency is the lynchpin of the Court’s decision—only by using this approach can the Court feel itself justified in declaring that “this is an extraordinary case” in an effort to limit both the known and unknown ramifications of its opinion.

It is the unknown ramifications of the Court’s new test that ought to be of the most concern to the Court. The Court’s new test will no doubt apply to delegations ranging from “school choice” to private prisons. In fact, “private” delegations are used extensively in Texas government, have been routinely upheld by this and other Texas courts, and given proper standards and oversight by the delegating authority, these delegations raise no constitutional concerns. For instance . . . , this Court [has] upheld the constitutionality of the Workers’ Compensation Act against [a] . . . challenge arising out of the use of the American Medical Association’s Guides to the Evaluation of Permanent Impairment for determining injured workers’ impairment ratings. Clearly, the Act delegates to the American Medical Association (AMA), a private national association of physicians, the task of promulgating standards for translating a worker’s job-related injury into an impairment rating to determine compensation benefits. . . . Yet the Court had no trouble upholding the use of the AMA Guides.

[Texas courts have also upheld] a legislative delegation to the Texas Automobile Insurance Plan Association. The Association, which is charged with administering the Safety-Responsibility Act’s scheme for providing motor vehicle liability insurance to high risk drivers, is composed exclusively of private insurance companies authorized to do business in Texas. Finding adequate safeguards against the arbitrary exercise of power under the statutory scheme, [a] court of appeals had no apparent trouble upholding this particular private delegation.

Nor does the Court consider the impact of its decision on the Legislature’s common practice of delegating eminent domain powers to private entities. Ironically, Justice Hecht
argues that the Foundation “wields more legislative power than any other privately chartered nonprofit corporation in Texas, or, as far as I can tell, in the history of Texas.” But not only have similar delegations to similar private organizations been upheld in other states, it is difficult for me to imagine a more profound delegation of governmental power than that given by the Texas Legislature to numerous private entities to take private property for public use under the sovereign power of eminent domain. How these delegations would fare under the Court’s new test is uncertain at best.

Even this Court delegates to a private entity the task of approving the law school course-of-study requirements for would-be Texas lawyers. . . . Not surprisingly, the Legislature has also embraced this model for a host of other certifications and licenses. Significantly, “private” delegations offer vast and varied expertise to Texas government at no additional cost to taxpayers and without bloating an already sizable bureaucracy.

Finally, only by leapfrogging a nearly unbroken string of decisions from this Court and the United States Supreme Court upholding legislative delegation of authority to various agencies, many with far less legislative guidance than was provided here, is the Court able to apply its new, stricter “private delegation” test in place of long-established Texas law. In my view, this Court’s decisions provide more than sufficient guidance—and certainly more authority—for evaluating the propriety of the delegation in this case. Our decisions also compel a determination that [the Act] is not facially unconstitutional.

II

Relying primarily on Louis Jaffe’s sixty-year-old law review article, the Court concludes that entities like the Foundation are wholly “private.” At the same time, the Court concedes that they do “not find it easy to categorize the Foundation as either a public or private agency.” The Court also fails to adequately explain why this largely fictional distinction, which leads it to propose an “either/or” choice, is so important that this entire statute should turn on it. I submit that the Foundation is neither wholly public nor wholly private, but rather it has attributes of each.

But as it turns out, this distinction does not really matter. What does matter more than drawing strained and artificial distinctions between public and private actors is whether the Foundation has been given adequate legislative guidance and whether its discretion is subject to appropriate oversight and other legal constraints. Ultimately, the legitimate constitutional concerns for unchecked abuse of power giving rise to the nondelegation doctrine are the same whether the Foundation is characterized as public or private. Thus, the Court’s creation and application of a new test for private delegations is simply unnecessary.

III

As I have previously noted, we have repeatedly held that the Legislature may delegate authority to agencies to carry out legislative purposes so long as it establishes “reasonable standards to guide the entity to which the powers are delegated.” “Such standards may be broad . . . [when] conditions must be considered which cannot be
conveniently investigated by the legislature,’” and at the same time, they should be “reasonably clear and hence acceptable as a standard of measurement.” The Court has recognized six categories of allowable delegations:

(1) delegations when the Legislature cannot practically and efficiently exercise powers (such as determining rail rates, questions of public convenience and necessity, and granting permission to drill oil wells);

(2) delegations to administrative bodies of the authority to make rules to implement statutes;

(3) delegations to find facts and ascertain conditions upon which an existing law may operate (for example, the authority given the Railroad Commission, the Public Utility Commission, the Texas Natural Resource Conservation Commission);

(4) delegations of legislative authority that set up broad standards, leaving to the delegate to make rules and determine facts to which the legislative policy is to apply, when the conditions to be considered cannot be easily investigated by the Legislature;

(5) delegations of the power to fix rates within prescribed limits to cover specified items of cost; and

(6) delegations of the power to determine the question of necessity of taking land for public use.

In testing the reasonableness of a particular delegation, we have required varying degrees of specificity in standards depending on the nature of the power conferred, the recipient of the power, and the subject matter. As these cases demonstrate, we have upheld numerous delegations with far more general grants of authority than that in [the Act].

Notwithstanding the Court’s assertion that [the Act] lacks sufficiently specific standards, the statute contains relatively comprehensive standards to guide both the Commissioner of Agriculture (a statewide elected official) and the Foundation in their joint efforts to execute the legislative mandate to “suppress and eradicate boll weevils and other cotton pests.” Initially, the Commissioner is required to select a nonprofit organization that “can best carry out the purposes of [the Act]” and must certify that the organization’s petition “complies with [the Act]” and “can adequately represent the interests of cotton growers in the proposed eradication zones.” The Commissioner may revoke the certification if it “fails to meet the requirements of [the Act].” Apparently, this includes the power to revoke the Foundation’s certification for its failure to conduct or abide by the results of Board elections or assessment referenda pursuant to the requirements of [the Act], or its failure to obtain the Commissioner’s approval for changing the voting status of a Board member, or otherwise comply with the requirements of the statute. Additionally, the statute requires that the Board use revenue collected “solely to finance programs approved by the commissioner as consistent with [the Act] and applicable provisions of the constitution” (emphasis added).
The Court refuses to acknowledge that [the Act] confer[s] such broad oversight powers upon the Commissioner. [The Act] can, however, be reasonably read to give the Commissioner the power to revoke the Foundation’s authority to operate under certain circumstances. Similarly, the statute clearly requires Commissioner approval of Foundation expenditures. The procedures for implementing these duties are subject to the broad rulemaking power that the Legislature has conferred upon the Commissioner. . . . [T]he statute repeatedly requires the Foundation to operate in accordance with both statutory standards and the Commissioner’s explicit grants of oversight authority.

The Legislature has also specified the powers of the Board. These include the power to “conduct programs consistent with the declaration of policy” stated by the Legislature, to “borrow money as necessary to execute this chapter,” and to take other action and exercise other authority “as necessary to execute any act authorized by [the Act] or the Texas Non-Profit Corporation Act” (emphasis added). The statute also specifies when the Board may add an area to an eradication zone and when and how assessments can be collected in those added areas. In proposing assessments, the Foundation must determine the assessment “needed” to “finance programs of marketing, promotion, research, and education calculated to increase the production and use of cotton” and the Foundation may use the funds only for eradication in that zone, operating costs (including payments of debt), and “other programs consistent with the declaration of policy” as stated by the Legislature in the statute.

The statute also contains guidelines for how the Foundation should implement eradication programs. As the Court has noted, the statute requires the Foundation to use the “best available integrated pest management techniques.” What the Court fails to acknowledge, however, is that the phrase “integrated pest management” is expressly defined by statute as “the coordinated use of pest and environmental information with available pest control methods to prevent unacceptable levels of pest damage by the most economical means and with the least possible hazard to people, property, and the environment” (emphasis added). This definition surely contains sufficiently specific standards to pass muster under this Court’s nondelegation jurisprudence. Similarly, the definitions of “eradication” and “infested” contain additional specific guidance.

“‘Eradication’ means elimination of boll weevils or pink bollworms to the extent that the Commissioner does not consider further elimination of boll weevils or pink bollworms necessary to prevent economic loss to cotton growers” (emphasis added). “‘Infested’ means the presence of the boll weevil or pink bollworm in any life stage or the existence of generally accepted entomological evidence from which it may be concluded with reasonable certainty that the boll weevil or pink bollworm is present” (emphasis added).

Thus, the delegation to the Foundation fits the description of at least four types of allowable delegation . . . . First, the Legislature cannot practically and efficiently exercise many of the powers delegated to the Foundation, such as determining the level of assessments needed to achieve the statute’s goals. Second, the statute delegates to the Foundation and Commissioner the authority to make rules to implement the statute. Third, the Foundation and Commissioner are given the power to find facts and ascertain conditions upon which [the Act] may operate, such as determining when a field is infested and what crops present a threat to the program. Fourth, the statute sets up broad
standards, leaving to the delegate to make rules and determine facts to which the legislative policy is to apply, when the conditions to be considered, such as level of infestation and which types of integrated management practices to use, cannot be easily investigated by the Legislature.

IV

Not only does the Court wrongly apply the correct legal standard to decide the constitutionality of the statute, it reaches several erroneous conclusions even under its own new test. Though the Court purports to limit its new test to “private” delegations, it does not explain why our long-standing test for improper delegations is inadequate, nor does it explore the potential negative ramifications of this new standard to other delegations. If, as the Court recognizes, broad delegations of lawmaking authority without adequate safeguards are a threat to democratic government, what difference does it make whether the delegation is to a bureaucrat, an appointed commission, or a quasi-governmental entity like the Foundation? Given adequate standards and oversight, such as that imposed on the Foundation under [the Act], this delegation is both constitutionally permissible and in the public interest.

Though the Court’s decision does not ultimately rely on the Foundation’s authority to promulgate rules, the violation of which constitutes a criminal offense . . . the Court asserts that this authority “strongly suggests an improper private delegation.” I disagree. The prospect of invalidating a delegation on this basis jeopardizes innumerable other statutes covering a multitude of subjects in which the Legislature has directed that a violation of agency rules is a criminal offense. These grants of authority to promulgate rules punishable by a fine are no more expansive than necessary to allow the enforcement of an agency’s rules. In this case, why would the Legislature delegate authority to administer a boll weevil eradication program and yet withhold from the Foundation the power to enforce its rules? Of course, no growers complain that they have been prosecuted for violating the rules of the Foundation. And, although I would not strike down the statute on this basis, even the Court recognizes that the proper course of action would be to strike down the particular penalty provisions, and sever them from the rest of the statute. If this is so, then the fifth factor of the Court’s test is simply unnecessary.

That the delegation is to a quasi-governmental entity as opposed to a “pure” administrative agency is irrelevant when, as in this case, the entity is required to promulgate rules consistent with the statute conferring its authority and consistent with rules promulgated by the Commissioner. The Board is given the authority to set penalties, after reasonable notice, for a grower’s failure to pay assessments. But it is the statute itself, and not the Board, that determines the circumstances under which a grower’s crop may be destroyed for nonpayment of assessments and the specific procedures that must be followed before destroying the crop. Only the Commissioner, and not the Foundation, is given the authority to establish rules regarding when planting of cotton may be prohibited “if there is reason to believe planting will jeopardize the success of the program or present a hazard to public health or safety,” and to order destruction of cotton when such rules are violated. Only the Commissioner, and not the Foundation, is given the authority to destroy and treat volunteer or noncommercial cotton and to determine
when destruction of commercial cotton is “necessary to carry out the purposes of [the Act].” Notably, the statute requires that the initial zone referendum include a summary of the penalties for noncompliance with rules adopted under the statute. The growers, therefore, may vote down a zone referendum if they consider the penalties to be too severe. Notwithstanding the “David v. Goliath” characterization by Chief Justice Phillips and Justice Hecht, each cotton grower, no matter how large or how small, is eligible to vote in any election. In other words, one grower, one vote. Under these circumstances, the danger of large-grower domination in an election is minimal.

While the Court argues that there is a significant difference between the requirement that the Commissioner promulgate rules in some instances and the authority of the Commissioner to promulgate rules in other instances, as I have already noted, we have long held that it is not necessary for an administrative agency to establish detailed rules before carrying out its statutory duty when the statute provides a sufficiently definite standard. The Commissioner is himself politically accountable, and is certainly capable of promulgating additional rules consistent with the statute as he deems necessary. Even so, in those instances in which the statute allows, but does not require the Commissioner to promulgate rules, the statute provides sufficiently definite standards to pass constitutional muster.

The Court also errs in its conclusion that [the Act] lacks adequate procedures for the growers to contest Foundation actions. Rather, the cotton growers are provided with ample opportunity to challenge Foundation and Commissioner decisions and to seek redress in the courts when necessary. The statute itself repeatedly requires that notice be provided to those subject to its jurisdiction. The statute also subjects any rulemaking by the Commissioner and the Foundation to the general requirements of Texas law governing administrative procedure. This includes judicial review of such rules if it is alleged that a rule or its threatened application interferes with or impairs, or threatens to interfere with or impair, a legal right or privilege. In addition, any person against whom the Department of Agriculture has assessed a penalty is entitled to an appeal, including a hearing in accordance with the Texas Administrative Procedure Act. The Foundation is also required to appoint an assessment appeals committee to adopt policies and procedures for grower appeals of the amount of assessments. Supplementing these administrative safeguards, however, Texas law has long recognized that even in the absence of an express right to judicial review of agency actions, such a right is implied when the complainant alleges that the statute has been unconstitutionally applied. Plainly, meaningful administrative and judicial review is available to the growers.

Finally, the Court erroneously relies on violations of the statute to demonstrate the statute’s lack of safeguards, reasoning that since the statute did not prevent what happened, the safeguards are insufficient. That the Commissioner or the Foundation could violate [the Act], however, does not render the statute unconstitutional, especially under a facial challenge. Ultimately, the Court’s reasoning “defeat[s] the purpose of delegating legislative authority” by “[r]equiring the legislature to include every detail and anticipate unforeseen circumstances.”
As previously noted, the United States Supreme Court has used the delegation doctrine to strike down a statute only three times in our nation’s history: *Carter Coal*, *Schechter Poultry*, and *Panama Refining*. All three cases were decided over sixty years ago, during the heyday of the Court’s hostility to President Franklin Roosevelt’s New Deal. Although this Court has stricken statutes based on improper delegation in a few instances, we have never done so under circumstances like those in this case.

The Supreme Court has repeatedly upheld delegations in which, as in this case, the ultimate decision-making authority is vested either with a government agency or with the regulated community through referenda. For example, in *Currin*, the Supreme Court upheld the Secretary of Agriculture’s authority to designate markets where tobacco could be bought and sold at auction upon receiving approval of two-thirds of the growers voting in a prescribed referendum. In so holding, the Supreme Court stated, “This is not a case where Congress has attempted to abdicate, or to transfer to others, the essential legislative functions with which it is vested by the Constitution” . . . .

That same year the Court also upheld a statute that required any order of the Secretary of Agriculture to be approved by interested producers before it could become effective. *United States v. Rock Royal Co-op.* (1939). The Court held that, whether or not a referendum is necessary, inasmuch as Congress could put the order into effect without approval of anyone, requiring such approval is not an invalid delegation.

The Supreme Court has always upheld statutes in which the regulated community has been delegated considerable authority so long as the statute contains “intelligible principles” . . . and when an administrative agency is vested with authority to oversee the private delegates’ exercise of power. [See *Sunshine Anthracite*.]

Similarly, in *J.W. Hampton*, the Court upheld Congress’ delegation to the President of the power to increase or decrease import duties to equalize the differences in costs of production in the United States and competing countries. The Court recognized that Congress may feel itself unable conveniently to determine exactly when its exercise of the legislative power should become effective, because [the need to

---

5 Whether *Carter Coal* is a nondelegation case at all is debatable. In *Mistretta*, the Supreme Court recognized *Schechter Poultry* and *Panama Refining* as the only two statutes it had ever stricken down for improper delegation. As a federal district court judge, Justice Scalia maintained that *Carter Coal* rested primarily upon a denial of substantive due process rights. In any event, its precedential value is dubious at best.

6 With regards to one market, one of the disputed orders in *Currin* was approved by a referendum in which only 25 percent of the growers cast a vote. Over ninety percent of those voting, however, approved the referendum. By contrast, voter turnout in the Board elections and assessment referenda in this case ranged from 35-85%, depending on the zone. In addition, the Court concludes that because the elections are limited to eligible growers, thus is not “popular,” thus supporting its conclusion that the delegation is private. This conclusion ignores the fact that under our form of government, the franchise is ordinarily limited to those subject to the governing body’s jurisdiction. For example, suffrage in elections for city council, commonly understood to be “popular elections,” are limited to city residents. Were this not the case, grave constitutional objections would no doubt arise.
exercise the power] depend[s] on future conditions, and it may leave the
determination of such time to the decision of an Executive, or, as often happens in
matters of state legislation, it may be left to a popular vote of the residents of a
district to be [affected] by the legislation. While in a sense one may say that such
residents are exercising legislative power, it is not an exact statement, because the
power has already been exercised legislatively . . . , the condition of its legislation
going into effect being made dependent by the legislature on the expression of the
voters of a certain district. . . .

If Congress shall lay down by legislative act an intelligible principle to which the
person or body authorized to fix such rates is directed to conform, such legislative
action is not [an unconstitutional] delegation of legislative power.

(Emphasis added.) The Supreme Court’s reasoning in these cases is particularly
applicable to the present case. Under the statute, the Foundation is allowed to take action
only upon the voters’ approval of the creation of the zone and the assessments. As in
Currin and Rock Royal, the people subject to these assessments have the ultimate veto
power. Inasmuch as the Legislature could impose a regulatory assessment to be collected
for purposes of boll weevil eradication, a requirement that assessments proposed by the
Foundation be approved through a referendum of growers cannot be an invalid
delegation. Nor is the delegation of the authority to establish the level of assessments to
the Foundation an improper delegation because the Legislature has provided for
sufficient oversight by the Commissioner of Agriculture and has provided reasonably
comprehensive guidance for the remaining discretionary authority of the Foundation,
particularly given the expertise of the Foundation Board members. The Court contends
that the statute “fails to meet the seventh factor” of its new test because there are no
guarantees that the Board members are “experts.” This is an audacious assertion indeed.
Perhaps the Legislature recognized that the cotton growers who earn their living by
producing cotton and who elect the Board members are the best judges of the
qualifications of the Board members they choose to represent them. . . .

VII

Accordingly, I would hold that [the Act] represents a constitutionally valid delegation of
power and that petitioners’ facial challenge fails. The statute’s “reasonable standards”
provide adequate guidance to the Commissioner and Foundation in implementing those
aspects of the statute that “cannot be conveniently investigated by the Legislature.”
Moreover, the statute provides adequate safeguards against the Foundation’s abuse of
power, most importantly the growers’ right to judicial review and their right to opt in and
out of the program. . . .

§ 1.10. General notes on the non-delegation doctrine

1. Non-delegation in the states. Note that the non-delegation doctrine, as described
in the federal cases, stems from the Vesting Clauses of the federal Constitution,
and more generally from the separation of powers at the federal level. Therefore, it is not binding on the states, which aren’t constitutionally required to have the same structure as the federal government (and which are probably entitled to scrap separation of powers altogether). Still, non-delegation cases can show up as due process cases under the federal Constitution (see the New Motor Vehicle Board case, discussed in note 9 of the Notes on the earlier federal cases), or as due process or non-delegation cases under state constitutions (see note 4 of the Notes on the earlier federal cases). In practice, all states do have separation of powers, and as Tex. Boll Weevil shows, various states have experimented with doctrines that are in various ways stricter than the federal doctrine. For instance, in Nebraska, the legislature can’t delegate the authority to define crimes.

2. The public-private distinction. In applying the bare non-delegation doctrine, it’s hard enough to figure out whether the scope of the delegate’s discretion is too broad. If, as in Tex. Boll Weevil, one wants to create a separate doctrine for private delegations, one has to additionally decide whether the delegate is public or private. The following section, on state action doctrine, shows how hard this can be in the context of privatization; but § V.C of Justice Phillips’s opinion already suggests the difficulties. Do you find the factors of publicness and privateness listed in that section convincing? Ultimately, the Court finds the Boll Weevil Foundation private because its members are “interested.” Note the parallels to Carter Coal, the due-process overtones of the inquiry, as well as the essentialism of treating government employees as presumptively disinterested. See note 5 in the Notes on the earlier federal cases, supra.

3. Other separation of powers doctrines. Taking into account the doctrine’s origin in the Vesting Clauses, the non-delegation doctrine is most naturally interpreted to bar Congress from giving up too much power—regardless of the identity of the delegate. On the other hand, the Appointments Clause is more directly concerned with the delegate himself: “[The President] shall nominate, and by and with the Advice and Consent of the Senate, shall appoint . . . Officers of the United States . . . : but the Congress may by Law vest the Appointment of such inferior Officers, as they may think proper, in the President alone, in the Courts of Law, or in the Heads of Departments.” Art. II, § 2, cl. 2. Thus, any officer except “inferior Officers” must have Presidential appointment and Senate confirmation. Thus, some commentators argue that current privatization practice, which gives significant authority to contractors solely appointed by the Executive Branch, violates the Appointments Clause. See VERKUIL, OUTSOURCING SOVEREIGNTY: WHY PRIVATIZATION OF GOVERNMENT FUNCTIONS THREATENS DEMOCRACY AND WHAT WE CAN DO ABOUT IT (2007).

4. The state action doctrine. In practice, the state action doctrine—the subject of the next section of these materials—is used far more often in imposing constitutional limits on privatization. The doctrinal difference between the two is that, in the privatization context, while the state action doctrine asks whether “ostensibly private actors should be considered public for constitutional purposes,” delegation doctrine “accepts their private status and asks instead
whether the Constitution prohibits governments from delegating certain powers to private actors.” Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1437 (2003). Metzger explains the interaction between the two doctrines:

“Rather than linking the possibility that a private delegation will lead to government authority evading constitutional constraints to the delegation’s constitutionality, this concern is addressed separately by determining whether the private entities involved should be considered state actors for constitutional purposes. *Flagg Bros. v. Brooks* (1978) is a prime example. The central issue in *Flagg Bros.* was presented as whether a warehouser’s sale of property pursuant to a state statute authorizing such ‘self-help’ for creditors constituted state action. If so, under the Court’s jurisprudence on creditor remedies, due process would require the government to review the private warehouser’s claims to property before it could be sold. Another way of understanding this due process claim, however, is to argue that the power to resolve disputes on a nonconsensual basis, even subject to later state court review, is one that government cannot delegate to private actors without preserving some opportunity for prior review. By holding that the private entity was not a state actor, the Court in essence rejected both arguments, although it considered the latter private delegation claim only implicitly, in assessing whether nonconsensual resolution of disputes was a public function.

“Of greatest importance, the Court has never considered how well state action doctrine functions as a surrogate for direct private delegation analysis. Indeed, whatever questions may exist about the contours of existing private delegation analysis, one seemingly indisputable point is that the Court does not view the constitutionality of a private delegation as turning on whether the private delegate is subject to constitutional constraints. In none of the cases is this question even raised. This failure to consider the essential dichotomy between public and private delegates is more than passing strange. While some private delegations may appear simply to be realignments of private parties’ rights, thereby making private delegates’ constitutional exemption unproblematic, many clearly involve grants of government power. *Friends of the Earth, Inc. v. Laidlaw Environmental Services, Inc.* (2000), for example, involved delegation of the power to decide to seek civil penalties on behalf of the United States for violations of law. Private delegates’ exemption from constitutional constraints means that they can wield these government powers in ways that raise serious abuse of power concerns. Imagine, for example, an individual who commences a meritless suit for civil penalties against a company out of spite or because its owners are African American. Given this danger, it is remarkable that the Court has never seen fit to link the constitutionality of private delegations to the constitutional status of the delegate.” *Id.* at 1444–45.
 § 2. The State Action Doctrine

1. All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside. No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

5. The Congress shall have power to enforce, by appropriate legislation, the provisions of this article.

U.S. CONST., amend. XIV (1868)

Every person who, under color of any statute, ordinance, regulation, custom, or usage, of any State . . . , subjects, or causes to be subjected, any . . . person . . . to the deprivation of any rights, privileges, or immunities secured by the Constitution and laws, shall be liable to the party injured.

42 U.S.C. § 1983 (originally enacted in 1871)

§ 2.1. The Civil Rights Cases, 109 U.S. 3 (1883)

MR. JUSTICE BRADLEY delivered the opinion of the court.

[These cases were brought under the Civil Rights Act of 1875 by “persons of color” who were denied the accommodations of hotels, theatres, and trains. The statute provides:]

“SEC. 1. That all persons within the jurisdiction of the United States shall be entitled to the full and equal enjoyment of the accommodations, advantages, facilities, and privileges of inns, public conveyances on land or water, theatres, and other places of public amusement; subject only to the conditions and limitations established by law, and applicable alike to citizens of every race and color, regardless of any previous condition of servitude.”

[Is this] constitutional? . . . [I]t is the purpose of the law to declare that, in the enjoyment of the [these] accommodations and privileges . . . , no distinction shall be made between citizens of different race or color, or between those who have, and those who have not, been slaves. Its effect is to declare, that . . . colored citizens . . . and citizens of other races, shall have the same accommodations and privileges in all inns, public conveyances, and places of amusement, as are enjoyed by white citizens; and vice versa.

. . .
Has Congress constitutional power to make such a law? Of course, no one will contend that the power to pass it was contained in the Constitution before the adoption of the last three amendments. The power is sought, first, in the Fourteenth Amendment, and the views and arguments of distinguished Senators, advanced whilst the law was under consideration, claiming authority to pass it by virtue of that amendment, are the principal arguments adduced in favor of the power.

The first section of the Fourteenth Amendment (which is the one relied on), after declaring who shall be citizens of the United States, and of the several States, is prohibitory upon the States. It declares that:

“No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.”

It is State action of a particular character that is prohibited. Individual invasion of individual rights is not the subject-matter of the amendment. It has a deeper and broader scope. It nullifies and makes void all State legislation, and State action of every kind, which impairs the privileges and immunities of citizens of the United States, or which injures them in life, liberty or property without due process of law, or which denies to any of them the equal protection of the laws. It not only does this, but, in order that the national will, thus declared, may not be a mere brutum fulmen, the last section of the amendment invests Congress with power to enforce it by appropriate legislation. To enforce what? To enforce the prohibition. To adopt appropriate legislation for correcting the effects of such prohibited State laws and State acts, and thus to render them effectually null, void, and innocuous.

This is the legislative power conferred upon Congress, and this is the whole of it. It does not invest Congress with power to legislate upon subjects which are within the domain of State legislation; but to provide modes of relief against State legislation, or State action, of the kind referred to. It does not authorize Congress to create a code of municipal law for the regulation of private rights; but to provide modes of redress against the operation of State laws, and the action of State officers executive or judicial, when these are subversive of the fundamental rights specified in the amendment. Positive rights and privileges are undoubtedly secured by the Fourteenth Amendment; but they are secured by way of prohibition against State laws and State proceedings affecting those rights and privileges, and by power given to Congress to legislate for the purpose of carrying such prohibition into effect: and such legislation must necessarily be predicated upon such supposed State laws or State proceedings, and be directed to the correction of their operation and effect.

In this connection it is proper to state that civil rights, such as are guaranteed by the Constitution against State aggression, cannot be impaired by the wrongful acts of individuals, unsupported by State authority in the shape of laws, customs, or judicial or executive proceedings. The wrongful act of an individual, unsupported by any such authority, is simply a private wrong, or a crime of that individual; an invasion of the
rights of the injured party, it is true, whether they affect his person, his property, or his reputation; but if not sanctioned in some way by the State, or not done under State authority, his rights remain in full force, and may presumably be vindicated by resort to the laws of the State for redress. An individual cannot deprive a man of his right to vote, to hold property, to buy and to sell, to sue in the courts, or to be a witness or a juror; he may, by force or fraud, interfere with the enjoyment of the right in a particular case; he may commit an assault against the person, or commit murder, or use ruffian violence at the polls, or slander the good name of a fellow citizen; but, unless protected in these wrongful acts by some shield of State law or State authority, he cannot destroy or injure the right; he will only render himself amenable to satisfaction or punishment; and amenable therefor to the laws of the State where the wrongful acts are committed. Hence, in all those cases where the Constitution seeks to protect the rights of the citizen against discriminative and unjust laws of the State by prohibiting such laws, it is not individual offences, but abrogation and denial of rights, which it denounces, and for which it clothes the Congress with power to provide a remedy. This abrogation and denial of rights, for which the States alone were or could be responsible, was the great seminal and fundamental wrong which was intended to be remedied. And the remedy to be provided must necessarily be predicated upon that wrong. It must assume that in the cases provided for, the evil or wrong actually committed rests upon some State law or State authority for its excuse and perpetration. . . .

[The Court also rejects the argument that congressional authority for this statute can be found in the power to enforce the Thirteenth Amendment, abolishing slavery, because discrimination in hotels, trains, and theatres is too far removed from slavery.]

On the whole we are of opinion, that no countenance of authority for the passage of the law in question can be found in . . . the . . . Fourteenth Amendment . . . .

MR. JUSTICE HARLAN, dissenting.

. . . [Justice Harlan first examines the Thirteenth Amendment argument.] [T]he power of Congress under the Thirteenth Amendment is not necessarily restricted to legislation against slavery as an institution upheld by positive law, but may be exerted to the extent, at least, of protecting the liberated race against discrimination, in respect of legal rights belonging to freemen, where such discrimination is based upon race.

It remains now to inquire what are the legal rights of colored persons in respect of the accommodations, privileges and facilities of public conveyances, inns, and places of public amusement?

First, as to public conveyances on land and water. . . . [T]his court . . . [has] said that a common carrier is “in the exercise of a sort of public office, and has public duties to perform, from which he should not be permitted to exonerate himself without the assent of the parties concerned.” . . . [In another case] it was ruled that railroads are public highways, established, by authority of the State for the public use; that they are none the less public highways, because controlled and owned by private corporations; that it is a part of the function of government to make and maintain highways for the conveyance of
the public; that no matter who is the agent, and what is the agency, the function performed is *that of the State*; that although the owners may be private companies, they may be compelled to permit the public to use these works in the manner in which they can be used . . . .

Such being the relations these corporations hold to the public, it would seem that the right of a colored person to use an improved public highway, upon the terms accorded to freemen of other races, is as fundamental, in the state of freedom established in this country, as are any of the rights which my brethren concede to be so far fundamental as to be deemed the essence of civil freedom. . . .

*Second*, as to inns. The same general observations which have been made as to railroads are applicable to inns. The word ‘inn’ has a technical legal signification. It means, in the act of 1875, just what it meant at common law. A mere private boarding-house is not an inn, nor is its keeper subject to the responsibilities, or entitled to the privileges of a common innkeeper. “To constitute one an innkeeper, within the legal force of that term, he must keep a house of entertainment or lodging for all travellers or wayfarers who might choose to accept the same, being of good character or conduct.” . . .

[Justice Harlan cites a treatise by Justice Story, and also a British case that states that] “innkeepers are a sort of public servants . . . .”

These authorities are sufficient to show a keeper of an inn is in the exercise of a quasi public employment. The law gives him special privileges, and he is charged with certain duties and responsibilities to the public. The public nature of his employment forbids him from discriminating against any person asking admission as a guest on account of the race or color of that person.

*Third*. As to places of public amusement. It may be argued that the managers of such places have no duties to perform with which the public are, in any legal sense, concerned, or with which the public have any right to interfere; and, that the exclusion of a black man from a place of public amusement, on account of his race, or the denial to him, on that ground, of equal accommodations at such places, violates no legal right for the vindication of which he may invoke the aid of the courts. My answer is, that places of public amusement, within the meaning of the act of 1875, are such as are established and maintained under direct license of the law. The authority to establish and maintain them comes from the public. The colored race is a part of that public. The local government granting the license represents them as well as all other races within its jurisdiction. A license from the public to establish a place of public amusement, imports, in law, equality of right, at such places, among all the members of that public. This must be so, unless it be—which I deny—that the common municipal government of all the people may, in the exertion of its powers, conferred for the benefit of all, discriminate or authorize discrimination against a particular race, solely because of its former condition of servitude.

I also submit whether it can be said—in view of the doctrines of this court as announced in *Munn v. Illinois* (1876) . . . —that the management of places of public amusement is a
purely private matter, with which government has no rightful concern? In the Munn case the question was whether the State of Illinois could fix, by law, the maximum of charges for the storage of grain in certain warehouses in that state—the private property of individual citizens. After quoting a remark attributed to Lord Chief Justice Hale, to the effect that when private property is “affected with a public interest it ceases to be juris privati only,” the court says:

“Property does become clothed with a public interest when used in a manner to make it of public consequence and affect the community at large. . . .”

. . . I am justified, upon the authority of that case, in saying that places of public amusement, conducted under the authority of the law, are clothed with a public interest, because used in a manner to make them of public consequence and to affect the community at large. The law may therefore regulate, to some extent, the mode in which they shall be conducted, and, consequently, the public have rights in respect of such places, which may be vindicated by the law. It is consequently not a matter purely of private concern. . . .

[Therefore these types of discrimination are badges of servitude that Congress may prevent under its Thirteenth Amendment enforcement power. Now let us consider the Fourteenth Amendment argument.] . . .

[Previous cases have held], and my brethren concede, that positive rights and privileges were intended to be secured, and are in fact secured, by the Fourteenth Amendment. . . .

The assumption that this amendment consists wholly of prohibitions upon State laws and State proceedings in hostility to its provisions, is unauthorized by its language. The first clause of the first section—“All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States, and of the State wherein they reside”—is of a distinctly affirmative character. . . . [It brought former slaves] within the direct operation of that provision of the Constitution which declares that “the citizens of each State shall be entitled to all privileges and immunities of citizens in the several States.” Art. 4, § 2.

The citizenship thus acquired, by that race, in virtue of an affirmative grant from the nation, may be protected, not alone by the judicial branch of the government, but by congressional legislation of a primary direct character; this, because the power of Congress is not restricted to the enforcement of prohibitions upon State laws or State action. It is, in terms distinct and positive, to enforce “the provisions of this article” of amendment; not simply those of a prohibitive character, but the provisions—all of the provisions—affirmative and prohibitive, of the amendment. It is, therefore, a grave misconception to suppose that the fifth section of the amendment has reference exclusively to express prohibitions upon State laws or State action. If any right was created by that amendment, the grant of power, through appropriate legislation, to enforce its provisions, authorizes Congress, by means of legislation, operating throughout the entire Union, to guard, secure, and protect that right.
It is, therefore, an essential inquiry what, if any, right, privilege or immunity was given, by the nation, to colored persons, when they were made citizens of the State in which they reside? . . . To this it may be answered, generally . . . , that they are those which are fundamental in citizenship in a free republican government, “common to the citizens in the latter States under their constitutions and laws by virtue of their being citizens.” . . .

. . . There is one, if there be no other—exemption from race discrimination in respect of any civil right belonging to citizens of the white race in the same State. That, surely, is their constitutional privilege when within the jurisdiction of other States. And such must be their constitutional right, in their own State, unless the recent amendments be splendid baubles, thrown out to delude those who deserved fair and generous treatment at the hands of the nation. Citizenship in this country necessarily imports at least equality of civil rights among citizens of every race in the same State. It is fundamental in American citizenship that, in respect of such rights, there shall be no discrimination by the State, or its officers, or by individuals or corporations exercising public functions or authority, against any citizen because of his race . . . .

[But even conceding that State action is required], I maintain that the decision of the court is erroneous. There has been adverse State action within the Fourteenth Amendment as heretofore interpreted by this court. I allude to Ex parte Virginia. It appears, in that case, that one Cole, judge of a county court, was charged with the duty, by the laws of Virginia, of selecting grand and petit jurors. The law of the State did not authorize or permit him, in making such selections, to discriminate against colored citizens because of their race. But he was indicted in the federal court, under the act . . . , for making such discriminations. The attorney-general of Virginia contended before us, that the State had done its duty, and had not authorized or directed that county judge to do what he was charged with having done; that the State had not denied to the colored race the equal protection of the laws; and that consequently the act of Cole must be deemed his individual act, in contravention of the will of the State. Plausible as this argument was, it failed to convince this court, and after saying that the Fourteenth Amendment had reference to the political body denominated a State, “by whatever instruments or in whatever modes that action may be taken,” and that a State acts by its legislative, executive, and judicial authorities, and can act in no other way, we proceeded:

“The constitutional provision, therefore, must mean that no agency of the State, or of the officers or agents by whom its powers are exerted, shall deny to any person within its jurisdiction the equal protection of the laws. . . .”

In every material sense applicable to the practical enforcement of the Fourteenth Amendment, railroad corporations, keepers of inns, and managers of places of public amusement are agents or instrumentalities of the State, because they are charged with duties to the public, and are amenable, in respect of their duties and functions, to governmental regulation. It seems to me that, within the principle settled in Ex parte Virginia, a denial, by these instrumentalities of the State, to the citizen, because of his race, of that equality of civil rights secured to him by law, is a denial by the State, within the meaning of the Fourteenth Amendment. If it be not, then that race is left, in respect of
the civil rights in question, practically at the mercy of corporations and individuals wielding power under the States.

But the court says that Congress did not . . . assume . . . to adjust what may be called the social rights of men and races in the community. I agree that government has nothing to do with social, as distinguished from technically legal, rights of individuals. No government ever has brought, or ever can bring, its people into social intercourse against their wishes. Whether one person will permit or maintain social relations with another is a matter with which government has no concern. I agree that if one citizen chooses not to hold social intercourse with another, he is not and cannot be made amenable to the law for his conduct in that regard; for even upon grounds of race, no legal right of a citizen is violated by the refusal of others to maintain merely social relations with him. What I affirm is that no State, nor the officers of any State, nor any corporation or individual wielding power under State authority for the public benefit or the public convenience, can, consistently either with the freedom established by the fundamental law, or with that equality of civil rights which now belongs to every citizen, discriminate against freemen or citizens, in those rights, because of their race, or because they once labored under the disabilities of slavery imposed upon them as a race. The rights which Congress . . . endeavored to secure and protect are legal, not social rights. The right, for instance, of a colored citizen to use the accommodations of a public highway, upon the same terms as are permitted to white citizens, is no more a social right than his right, under the law, to use the public streets of a city or a town, or a turnpike road, or a public market, or a post office, or his right to sit in a public building with others, of whatever race, for the purpose of hearing the political questions of the day discussed. Scarcely a day passes without our seeing in this court-room citizens of the white and black races sitting side by side, watching the progress of our business. It would never occur to any one that the presence of a colored citizen in a court-house, or court-room, was an invasion of the social rights of white persons who may frequent such places. And yet, such a suggestion would be quite as sound in law . . . as is the suggestion that the claim of a colored citizen to use, upon the same terms as is permitted to white citizens, the accommodations of public highways, or public inns, or places of public amusement, established under the license of the law, is an invasion of the social rights of the white race. . . .

§ 2.2. Shelley v. Kraemer, 334 U.S. 1 (1948)

MR. CHIEF JUSTICE VINSON delivered the opinion of the Court.

These cases present for our consideration questions relating to the validity of court enforcement of private agreements, generally described as restrictive covenants, which [exclude] persons of designated race or color from the ownership or occupancy of real property. . . .

[A group of property owners signed a deed restriction preventing occupancy by anyone but white people for the next 50 years. Later, black people (without knowledge of the deed restriction) bought the property. Another property owner subject to the covenant]
sued. The black owners argued that judicial enforcement of the covenant violated their Fourteenth Amendment rights.

I.

. . . It cannot be doubted that among the civil rights . . . protected . . . by the Fourteenth Amendment are the rights to acquire, enjoy, own and dispose of property. Equality in the enjoyment of property rights was regarded by the framers of that Amendment as an essential pre-condition to the realization of other basic civil rights and liberties which the Amendment was intended to guarantee. Thus, . . . § 1 of the Civil Rights Act of 1866 which was enacted by Congress while the Fourteenth Amendment was also under consideration, provides:

“All citizens of the United States shall have the same right . . . as is enjoyed by white citizens thereof to inherit, purchase, lease, sell, hold, and convey real and personal property.” . . .

It is likewise clear that [such] restrictions [would be unconstitutional] if imposed by state statute or local ordinance. . . .

But the present cases . . . do not involve action by state legislatures or city councils. Here the particular patterns of discrimination . . . are determined, in the first instance, by the terms of agreements among private individuals. Participation of the State consists in the enforcement of the restrictions so defined. The crucial issue with which we are here confronted is whether this distinction removes these cases from the operation of the prohibitory provisions of the Fourteenth Amendment.

Since . . . the Civil Rights Cases, the principle has become firmly embedded in our constitutional law that the action inhibited by the first section of the Fourteenth Amendment is only such action as may fairly be said to be that of the States. That Amendment erects no shield against merely private conduct, however discriminatory or wrongful.

We conclude, therefore, that the restrictive agreements standing alone cannot be regarded as violative of any [Fourteenth Amendment] rights . . . [s]o long as the purposes of those agreements are effectuated by voluntary adherence to their terms . . . .

But here there was more. These are cases in which the purposes of the agreements were secured only by judicial enforcement by state courts . . . . The respondents urge that judicial enforcement of private agreements does not amount to state action; or, in any event, the participation of the State is so attenuated in character as not to amount to state action within the meaning of the Fourteenth Amendment. Finally, it is suggested, even if the States in these cases may be deemed to have acted in the constitutional sense, their action did not deprive petitioners of rights guaranteed by the Fourteenth Amendment. . . .
II.

That the action of state courts and judicial officers in their official capacities is to be regarded as action of the State within the meaning of the Fourteenth Amendment, is a proposition which has long been established by decisions of this Court. . . .

[For instance, unfair judicial proceedings, like the exclusion of Negroes from jury service, or denial of notice and opportunity to defend, have long been held to violate the Fourteenth Amendment.] . . .

But [in addition,] . . . the action of state courts in enforcing a substantive common-law rule formulated by those courts, may result in the denial of rights guaranteed by the Fourteenth Amendment, even though the judicial proceedings in such cases may have been in complete accord with the most rigorous conceptions of procedural due process. Thus . . ., enforcement by state courts of the common-law policy of the State, which resulted in the restraining of peaceful picketing, was held to be state action of the sort prohibited by the Amendment’s guaranties of freedom of discussion . . . [A] conviction in a state court of the common-law crime of breach of the peace was, under the circumstances of the case, found to be a violation of the Amendment’s commands relating to freedom of religion . . . [E]nforcement of the state’s common-law rule relating to contempts by publication was held to be state action inconsistent with the prohibitions of the Fourteenth Amendment.

. . . [I]t has never been suggested that state court action is immunized from the operation of those provisions simply because the act is that of the judicial branch of the state government.

III.

. . . We have no doubt that there has been state action in these cases in the full and complete sense of the phrase. . . . [P]etitioners were willing purchasers of properties upon which they desired to establish homes. The owners of the properties were willing sellers; and contracts of sale were accordingly consummated. It is clear that but for the active intervention of the state courts, supported by the full panoply of state power, petitioners would have been free to occupy the properties in question without restraint.

These are not cases, as has been suggested, in which the States have merely abstained from action, leaving private individuals free to impose such discriminations as they see fit. Rather, these are cases in which the States have made available to such individuals the full coercive power of government to deny to petitioners, on the grounds of race or color, the enjoyment of property rights in premises which petitioners are willing and financially able to acquire and which the grantors are willing to sell. The difference between judicial enforcement and nonenforcement of the restrictive covenants is the difference to petitioners between being denied rights of property available to other members of the community and being accorded full enjoyment of those rights on an equal footing.

The enforcement of the restrictive agreements by the state courts in these cases was directed pursuant to the common-law policy of the States . . . . The judicial action in each
case bears the clear and unmistakable imprimatur of the State. . . . Nor is the Amendment ineffective simply because the particular pattern of discrimination, which the State has enforced, was defined initially by the terms of a private agreement. State action, as that phrase is understood for the purposes of the Fourteenth Amendment, refers to exertions of state power in all forms. . . .

We hold that in granting judicial enforcement of the restrictive agreements in these cases, the States have denied petitioners the equal protection of the laws . . . . [F]reedom from discrimination by the States in the enjoyment of property rights was among the basic objectives sought to be effectuated by the framers of the Fourteenth Amendment. That such discrimination has occurred in these cases is clear. Because of the race or color of these petitioners they have been denied rights of ownership or occupancy enjoyed as a matter of course by other citizens of different race or color. . . .

Respondents urge, however, that since the state courts stand ready to enforce restrictive covenants excluding white persons from the ownership or occupancy of property covered by such agreements, enforcement of covenants excluding colored persons may not be deemed a denial of equal protection of the laws to the colored persons who are thereby affected. This contention does not bear scrutiny. The parties have directed our attention to no case in which a court, state or federal, has been called upon to enforce a covenant excluding members of the white majority from ownership or occupancy of real property on grounds of race or color. But there are more fundamental considerations. The rights created by the first section of the Fourteenth Amendment are, by its terms, guaranteed to the individual. The rights established are personal rights. It is, therefore, no answer to these petitioners to say that the courts may also be induced to deny white persons rights of ownership and occupancy on grounds of race or color. Equal protection of the laws is not achieved through indiscriminate imposition of inequalities.

§ 2.3. Notes on the early state-action cases

1. State-action doctrine as “transsubstantive.” Several constitutional doctrines have a state-action requirement. As three examples, consider the Due Process Clause (implicated in Lugar and Sullivan), the Equal Protection Clause (implicated in Shelley), and the Free Speech Clause (implicated in Rendell-Baker and Brentwood Academy). The state-action doctrine is applied equivalently across all these fields; in other words, there’s nothing you need to know about First Amendment or Equal Protection law to apply the doctrine. Some commentators think the entire limitation of constitutional rights to state action is wrong-headed. See, e.g., Chemerinsky, Rethinking State Action, 80 NW. L. REV. 503 (1985); see also the Notes on the conceptual difficulties of the state action doctrine, infra. But suppose you agree with the concept of state action in principle. Is there any reason why one might oppose transsubstantiveness—that is, why one might think some actor is the State for purposes of the Free Speech Clause but not for purposes of the Equal Protection Clause?
2. **What is not state action? The Harlan dissent.** Consider Justice Harlan’s dissent in *The Civil Rights Cases*. He explains that the Fourteenth Amendment protects rights “fundamental in citizenship,” which includes freedom from “race discrimination in respect of any civil right,” which applies against the State or “individuals or corporations exercising public functions or authority.” This raises the obvious question: why are hotels, trains, and theatres “exercising public functions or authority,” rather than being (as the majority says) merely private entities?

Justice Harlan gives the answer in his preceding discussion of the Thirteenth Amendment argument, where he discusses the legal rights of freemen. As to public conveyances and inns, he discusses the traditional common-law rule that such establishments were required to serve all comers, quoting legal sources characterizing them as “a sort of public office” (performing the function “of the State”) or “a sort of public servant” (with a “quasi public employment”). If this were all, Harlan’s theory would be limited to particular special professions that could be considered quasi-governmental.

But consider his justification as regards “places of public amusement”: such places operate under licenses, are “clothed with a public interest” because they “affect the community at large,” and may therefore be regulated. (In Harlan’s alternative formulation, all three of these categories are “agents or instrumentalities of the State” because they are “amenable . . . to government regulation.”) Especially after the New Deal-era constitutional revolution, is any area immune from government regulation? Would Harlan’s theory leave anything untouched by the Fourteenth Amendment?

3. **What is not state action? Shelley v. Kraemer.** Consider the Court’s opinion in *Shelley*: (1) The property owners’ discriminatory covenants are not state action. (2) The action of courts, however, is state action, even when they do nothing but enforce common-law rules (such as rules against picketing or libel, which may violate First Amendment guarantees). (3) Thus, the courts’ enforcement of the discriminatory covenant violates the Equal Protection Clause because it “ma[kes] available . . . the full coercive power of government to deny . . . on the grounds of race . . . the enjoyment of property rights.” Can one not agree with (1) and (2) without signing on to (3)? The state’s policy can be characterized as one of evenhandedly enforcing contracts, including restrictive covenants (regardless whether they are racially restrictive). Thus, there is discrimination, but (per (1)) it is not state action; and there is state action (per (2)), but it is nondiscriminatory. Under the holding in *Shelley*, what is not state action?

Thus, though there was no dissent in *Shelley*, and though the result in that case seems attractive, *Shelley* has turned out to be quite difficult to explain theoretically in ways consistent with the state-action doctrine. Scholarly commentary has either (like Chemerinsky) sought to reject the public-private distinction, argued that the state common law in *Shelley* delegated the zoning power to private parties, interpreted *Shelley* as introducing a balancing test, or

Shelley could have been decided exclusively based on 42 U.S.C. § 1982, quoted in the opinion as § 1 of the Civil Rights Act of 1866. This statute is now considered to be grounded in the Thirteenth Amendment enforcement power (on the theory that race discrimination in the right to make contracts or hold property is an incident of slavery), though this was not the case in 1948. See Hurd v. Hodge (1948). The Court did not interpret § 1982 as covering private action until Jones v. Alfred H. Mayer Co. in 1968. (One can also argue that such covenants violate § 1981, which guarantees equal rights to “make and enforce contracts,” but that statute was not held to cover private action until Runyon v. McCrary in 1976.)

4. **Burton v. Wilmington Parking Auth. (1961) and “joint participation.”** A black man was denied service in a restaurant that leased space in a municipal parking structure. The Parking Authority put the space out to bid (and ultimately leased it out to the restaurant and other enterprises) because the structure would not be financially sustainable based only on parking revenues. The Supreme Court, in an opinion by Justice Clark, held that the discrimination was state action. The majority opinion stressed the intensely fact-bound nature of this determination, but one of the main considerations was that “the peculiar relationship of the restaurant to the parking facility in which it is located confers on each an incidental variety of mutual benefits”: restaurant guests get parking while the city gets parking revenue; and any improvements made by the restaurant would not increase its taxes because the fee is held by a tax-exempt government agency. “Neither can it be ignored, especially in view of Eagle’s affirmative allegation that for it to serve Negroes would injure its business, that profits earned by discrimination not only contribute to, but also are indispensable elements in, the financial success of a governmental agency.” This and some other factors made for an unacceptable degree of “state participation and involvement in discriminatory action.” The Court also noted the irony that discrimination was illegal in the parking structure while it was legal in the city’s lessee’s restaurant. “The State has so far insinuated itself into a position of interdependence with Eagle that it must be recognized as a joint participant in the challenged activity, which, on that account, cannot be considered to have been so ‘purely private’ as to fall without the scope of the Fourteenth Amendment.”

Assume that a private firm generally engages in behavior because it believes that behavior to be profit-maximizing. (This may not always be true of discrimination, but much discrimination may in fact be profit-maximizing to the extent it caters to the discriminatory views of the firm’s clientele.) Because governments generally benefit from the profits of firms within its jurisdiction—not only by increased tax revenue, but also by increased employment and the like—don’t profits earned by any private discrimination always contribute to the financial success of a
government? Does it make a difference that the restaurant’s rent was a large part of the revenues of the parking structure? If so, would the answer change if the parking structure were managed directly by the state of Delaware together with all other state properties? Imagine the city of Wilmington owned a general-purpose office building, and its tenants include Democratic and Republican party offices, which hire only employees that agree with their party platforms; a Christian seminary, which hires only employees that believe its brand of Christianity; a firm owned by a patriotic American who will only hire citizens; and various retail stores that practice at-will employment (i.e., without any “due process” protections). Assume that all these practices would be constitutional if the tenants were at the private office building across the street. How would Burton apply?

Note that the majority and dissents in Blum spar over Burton, and Burton is also mentioned in Rendell-Baker and in Brennan’s dissent in DeShaney. In Sullivan, the Court describes it as an “early” state-action opinion whose test has been “refined” by later cases, which suggests that it should be read very narrowly.


MR. JUSTICE REHNQUIST delivered the opinion of the Court.

The question presented by this litigation is whether a warehouseman’s proposed sale of goods entrusted to him for storage, as permitted by New York Uniform Commercial Code § 7-210, is an action properly attributable to the State of New York. . . .

I

According to her complaint, the allegations of which we must accept as true, respondent Shirley Brooks and her family were evicted . . . . The city marshal arranged for Brooks’ possessions to be stored by petitioner Flagg Brothers, Inc., in its warehouse. Brooks was informed of the cost of moving and storage, and she instructed the workmen to proceed, although she found the price too high. . . . [A]fter a series of disputes over the validity of the charges being claimed by petitioner Flagg Brothers, Brooks received a letter demanding that her account be brought up to date within 10 days “or your furniture will be sold.” . . .

Brooks thereupon initiated this class action . . . under 42 U.S.C. § 1983, seeking damages, an injunction against the threatened sale of her belongings, and the declaration that such a sale pursuant to § 7-210 would violate the Due Process and Equal Protection Clauses . . . . She was later joined . . . by Gloria Jones, another resident . . . whose goods had been stored by Flagg Brothers following her eviction. . . .
II

A claim . . . under § 1983 must embody at least two elements. Respondents are first bound to show that they have been deprived of a right “secured by the Constitution and the laws” of the United States. They must secondly show that Flagg Brothers deprived them of this right acting “under color of any statute” of the State of New York. It is clear that these two elements denote two separate areas of inquiry. . . .

Respondents allege in their complaints that “the threatened sale of the goods pursuant to New York Uniform Commercial Code § 7-210” is an action under color of state law. We have previously noted, with respect to a private individual, that “[w]hatever else may also be necessary to show that a person has acted ‘under color of [a] statute’ for purposes of § 1983, . . . we think it essential that he act with the knowledge of and pursuant to that statute.” Certainly, the complaints can be fairly read to allege such knowledge on the part of Flagg Brothers. However, we need not determine whether any further showing is necessary, since it is apparent that neither respondent has alleged facts which constitute a deprivation of any right “secured by the Constitution and laws” of the United States.

A moment’s reflection will clarify the essential distinction between the two elements of a § 1983 action. . . . [M]ost rights secured by the Constitution are protected only against infringement by governments. . . . Here, respondents allege that Flagg Brothers has deprived them of their right, secured by the Fourteenth Amendment, to be free from state deprivations of property without due process of law. Thus, they must establish not only that Flagg Brothers acted under color of the challenged statute, but also that its actions are properly attributable to the State of New York. [Of course, where the defendant is a public official, the two elements of a § 1983 action merge.]

It must be noted that respondents have named no public officials as defendants in this action. The City Marshal, who supervised their evictions, was dismissed from the case by the consent of all the parties. This total absence of overt official involvement plainly distinguishes this case from earlier decisions imposing procedural restrictions on creditors’ remedies such as North Georgia Finishing, Inc. v. Di-Chem, Inc. (1975); Fuentes v. Shevin (1972); Sniadach v. Family Finance Corp. (1969). In those cases, the Court was careful to point out that the dictates of the Due Process Clause “attach[h] only to the deprivation of an interest encompassed within the Fourteenth Amendment’s protection.” While as a factual matter any person with sufficient physical power may deprive a person of his property, only a State or a private person whose action “may be fairly treated as that of the State itself” may deprive him of “an interest encompassed within the Fourteenth Amendment’s protection.” Thus, the only issue presented by this case is whether Flagg Brothers’ action may fairly be attributed to the State of New York. We conclude that it may not.

III

Respondents’ primary contention is that New York has delegated to Flagg Brothers a power “traditionally exclusively reserved to the State.” They argue that the resolution of private disputes is a traditional function of civil government, and that the State in § 7-210
has delegated this function to Flagg Brothers. Respondents, however, have read too much into the language of our previous cases. While many functions have been traditionally performed by governments, very few have been “exclusively reserved to the State.”

One such area has been elections. . . . A second line of cases under the public-function doctrine originated with *Marsh v. Alabama* (1946)[, where a private company performed all the necessary municipal functions in the town of Chickasaw, Ala., which it owned]. . . . These two branches of the public-function doctrine have in common the feature of exclusivity. . . .

[T]he proposed sale by Flagg Brothers under § 7-210[, however,] is not the only means of resolving this purely private dispute. Respondent Brooks has never alleged that state law barred her from seeking a waiver of Flagg Brothers’ right to sell her goods at the time she authorized their storage. Presumably, respondent Jones, who alleges that she never authorized the storage of her goods, could have sought to replevy her goods at any time under state law. The challenged statute itself provides a damages remedy against the warehouseman for violations of its provisions. This system of rights and remedies, recognizing the traditional place of private arrangements in ordering relationships in the commercial world, can hardly be said to have delegated to Flagg Brothers an exclusive prerogative of the sovereign.

Whatever the particular remedies available under New York law, we do not consider a more detailed description of them necessary to our conclusion that the settlement of

---

9 Unlike the parade of horrors suggested by our Brother STEVENS in dissent, this case does not involve state authorization of private breach of the peace.

10 It is undoubtedly true, as our Brother STEVENS says in dissent, that “respondents have a property interest in the possessions that the warehouseman proposes to sell.” But that property interest is not a monolithic, abstract concept hovering in the legal stratosphere. It is a bundle of rights in personality, the metes and bounds of which are determined by the decisional and statutory law of the State of New York. The validity of the property interest in these possessions which respondents previously acquired from some other private person depends on New York law, and the manner in which that same property interest in these same possessions may be lost or transferred to still another private person likewise depends on New York law. It would intolerably broaden, beyond the scope of any of our previous cases, the notion of state action under the Fourteenth Amendment to hold that the mere existence of a body of property law in a State, whether decisional or statutory, itself amounted to “state action” even though no process or state officials were ever involved in enforcing that body of law. This situation is clearly distinguishable from cases such as *Di-Chem*, *Fuentes*, and *Sniadach*. In each of those cases a government official participated in the physical deprivation of what had concededly been the constitutional plaintiff’s property under state law before the deprivation occurred. The constitutional protection attaches not because, as in *Di-Chem*, a clerk issued a ministerial writ out of the court, but because as a result of that writ the property of the debtor was seized and impounded by the affirmative command of the law of Georgia. The creditor in *Di-Chem* had not simply sought to pursue the collection of his debt by private means permissible under Georgia law; he had invoked the authority of the Georgia court, which in turn had ordered the garnishee not to pay over money which previously had been the property of the debtor. See *Shelley*. The “consent” inquiry in *Fuentes* occurred only after the Court had concluded that state action for purposes of the Fourteenth Amendment was supplied by the participation in the seizure on the part of the sheriff. The consent inquiry was directed to whether there had been a waiver of the constitutional right to due process which had been triggered by state deprivation of property. But our Brother STEVENS puts the cart before the horse; he concludes that the respondents’ lack of consent to the deprivations triggers affirmative constitutional protections which the State is bound to provide. Thus what was a mere coda to the constitutional analysis in *Fuentes* becomes the major theme of the dissent.
disputes between debtors and creditors is not traditionally an exclusive public function. 11

Thus, even if we were inclined to extend the sovereign-function doctrine outside of its present carefully confined bounds, the field of private commercial transactions would be a particularly inappropriate area into which to expand it. We conclude that our sovereign-function cases do not support a finding of state action here. . . .

IV

Respondents further urge that Flagg Brothers’ proposed action is properly attributable to the State because the State has authorized and encouraged it in enacting § 7-210. Our cases state “that a State is responsible for the . . . act of a private party when the State, by its law, has compelled the act.” This Court, however, has never held that a State’s mere acquiescence in a private action converts that action into that of the State. . . .

It is quite immaterial that the State has embodied its decision not to act in statutory form. If New York had no commercial statutes at all, its courts would still be faced with the decision whether to prohibit or to permit the sort of sale threatened here the first time an aggrieved bailor came before them for relief. A judicial decision to deny relief would be no less an “authorization” or “encouragement” of that sale than the legislature’s decision embodied in this statute. . . . If the mere denial of judicial relief is considered sufficient encouragement to make the State responsible for those private acts, all private deprivations of property would be converted into public acts whenever the State, for whatever reason, denies relief sought by the putative property owner. . . .

Here, the State of New York has not compelled the sale of a bailor’s goods, but has merely announced the circumstances under which its courts will not interfere with a private sale. Indeed, the crux of respondents’ complaint is not that the State has acted, but that it has refused to act. This statutory refusal to act is no different in principle from an ordinary statute of limitations whereby the State declines to provide a remedy for private deprivations of property after the passage of a given period of time.

11 It may well be, as my Brother STEVENS’ dissent contends, that “[t]he power to order legally binding surrenders of property and the constitutional restrictions on that power are necessary correlatives in our system.” But here New York, unlike Florida in Fuentes, Georgia in Di-Chem, and Wisconsin in Sniadach, has not ordered respondents to surrender any property whatever. It has merely enacted a statute which provides that a warehouseman conforming to the provisions of the statute may convert his traditional lien into good title. There is no reason whatever to believe that either Flagg Brothers or respondents could not, if they wished, seek resort to the New York courts in order to either compel or prevent the “surrenders of property” to which that dissent refers, and that the compliance of Flagg Brothers with applicable New York property law would be reviewed after customary notice and hearing in such a proceeding. The fact that such a judicial review of a self-help remedy is seldom encountered bears witness to the important part that such remedies have played in our system of property rights. This is particularly true of the warehouseman’s lien, which is the source of this provision in the Uniform Commercial Code which is the law in 49 States and the District of Columbia. The lien in this case, particularly because it is burdened by procedural constraints and provides for a compensatory remedy and judicial relief against abuse, is not atypical of creditors’ liens historically, whether created by statute or legislatively enacted. The conduct of private actors in relying on the rights established under these liens to resort to self-help remedies does not permit their conduct to be ascribed to the State.

108

Draft — do not distribute!
We conclude that the allegations of these complaints do not establish a violation of these respondents’ Fourteenth Amendment rights by either petitioner Flagg Brothers or the State of New York.

MR. JUSTICE STEVENS, with whom MR. JUSTICE WHITE and MR. JUSTICE MARSHALL join, dissenting.

Respondents contend that petitioner Flagg Brothers’ proposed sale of their property to third parties will violate the Due Process Clause of the Fourteenth Amendment. Assuming, arguendo, that the procedure to be followed would be inadequate if the sale were conducted by state officials, the Court holds that respondents have no federal protection because the case involves nothing more than a private deprivation of their property without due process of law. In my judgment the Court’s holding is fundamentally inconsistent with, if not foreclosed by, our prior decisions which have imposed procedural restrictions on the State’s authorization of certain creditors’ remedies.

There is no question in this case but that respondents have a property interest in the possessions that the warehouseman proposes to sell. It is also clear that, whatever power of sale the warehouseman has, it does not derive from the consent of the respondents. The claimed power derives solely from the State, and specifically from § 7-210 of the New York Uniform Commercial Code. The question is whether a state statute which authorizes a private party to deprive a person of his property without his consent must meet the requirements of the Due Process Clause of the Fourteenth Amendment. This question must be answered in the affirmative unless the State has virtually unlimited power to transfer interests in private property without any procedural protections.

In determining that New York’s statute cannot be scrutinized under the Due Process Clause, the Court reasons that the warehouseman’s proposed sale is solely private action because the state statute “permits but does not compel” the sale and because the warehouseman has not been delegated a power “exclusively reserved to the State.” Under this approach a State could enact laws authorizing private citizens to use self-help in countless situations without any possibility of federal challenge. A state statute could authorize the warehouseman to retain all proceeds of the lien sale, even if they far exceeded the amount of the alleged debt; it could authorize finance companies to enter private homes to repossess merchandise; or indeed, it could authorize “any person with

1 Of course the warehouseman may also have a property interest and the ultimate resolution of the due process issue will require a balancing of these interests.

2 Although the petitioners have at various stages of this case contended that there was an “implied contract” between the warehouseman and respondents providing for the sale of respondents’ possessions in satisfaction of a lien, the Court of Appeals rejected this claim, and petitioners conceded in this Court that, taking respondents’ allegations as fact, as we must, there is no contractual issue in this case.

3 It could be argued that since the State has the power to create property interests, it should also have the power to determine what procedures should attend the deprivation of those interests. See Arnett v. Kennedy (1974) (Rehnquist, J.). Although a majority of this Court has never adopted that position, today’s opinion revives the theory in a somewhat different setting by holding that the State can shield its legislation affecting property interests from due process scrutiny by delegating authority to private parties.
sufficient physical power” to acquire and sell the property of his weaker neighbor. An attempt to challenge the validity of any such outrageous statute would be defeated by the reasoning the Court uses today: The Court’s rationale would characterize action pursuant to such a statute as purely private action, which the State permits but does not compel, in an area not exclusively reserved to the State.

As these examples suggest, the distinctions between “permission” and “compulsion” on the one hand, and “exclusive” and “nonexclusive,” on the other, cannot be determinative factors in state-action analysis. There is no great chasm between “permission” and “compulsion” requiring particular state action to fall within one or the other definitional camp. . . . In this case, the State of New York, by enacting § 7-210 of the Uniform Commercial Code, has acted in the most effective and unambiguous way a State can act. This section specifically authorizes petitioner Flagg Brothers to sell respondents’ possessions; it details the procedures that petitioner must follow; and it grants petitioner the power to convey good title to goods that are now owned by respondents to a third party.

While Members of this Court have suggested that statutory authorization alone may be sufficient to establish state action, it is not necessary to rely on those suggestions in this case because New York has authorized the warehouseman to perform what is clearly a state function. The test of what is a state function for purposes of the Due Process Clause has been variously phrased. Most frequently the issue is presented in terms of whether the State has delegated a function traditionally and historically associated with sovereignty. In this Court, petitioners have attempted to argue that the nonconsensual transfer of property rights is not a traditional function of the sovereign. The overwhelming historical evidence is to the contrary, however, and the Court wisely does not adopt this position. Instead, the Court reasons that state action cannot be found because the State has not delegated to the warehouseman an exclusive sovereign function. This distinction, however, is not consistent with our prior decisions on state action; is not even adhered to by the Court in this case; and, most importantly, is inconsistent with the line of cases beginning with Sniadach.

Since Sniadach this Court has scrutinized various state statutes regulating the debtor-creditor relationship for compliance with the Due Process Clause. In each of these cases

---

8 As I understand the Court’s notion of “exclusivity,” the sovereign function here is not exclusive because there may be other state remedies, under different statutes or common-law theories, available to respondents. Even if I were to accept the notion that sovereign functions must be “exclusive,” the Court’s description of exclusivity is incomprehensible. The question is whether a particular action is a uniquely sovereign function, not whether state law forecloses any possibility of recovering for damages for such activity. For instance, it is clear that the maintenance of a police force is a unique sovereign function, and the delegation of police power to a private party will entail state action. Under the Court’s analysis, however, there would be no state action if the State provided a remedy, such as an action for wrongful imprisonment, for the individual injured by the “private” policeman. This analysis is not based on “exclusivity,” but on some vague, and highly inappropriate, notion that respondents should not complain about this state statute if the State offers them a glimmer of hope of redeeming their possessions, or at least the value of the goods, through some other state action. Of course, the availability of other state remedies may be relevant in determining whether the statute provides sufficient procedural protections under the Due Process Clause, but it is not relevant to the state-action issue.
[(Di-Chem, Mitchell v. W.T. Grant Co. (1974), and Fuentes)] a finding of state action was a prerequisite to the Court’s decision. The Court today seeks to explain these findings on the ground that in each case there was some element of “overt official involvement.” Given the facts of those cases, this explanation is baffling. In Di-Chem, for instance, the official involvement of the State of Georgia consisted of a court clerk who issued a writ of garnishment based solely on the affidavit of the creditor. The clerk’s actions were purely ministerial, and, until today, this court had never held that purely ministerial acts of “minor governmental functionaries” were sufficient to establish state action. The suggestion that this was the basis for due process review in Sniadach, Shevin, and Di-Chem marks a major and, in my judgment, unwise expansion of the state-action doctrine. The number of private actions in which a governmental functionary plays some ministerial role is legion[, for instance, in the transfer of cars or real estate]; to base due process review on the fortuity of such governmental intervention would demean the majestic purposes of the Due Process Clause.

Instead, cases such as Di-Chem must be viewed as reflecting this Court’s recognition of the significance of the State’s role in defining and controlling the debtor-creditor relationship. The Court’s language to this effect in the various debtor-creditor cases has been unequivocal. In Fuentes the Court stressed that the statutes in question “abdicate[d] effective state control over state power.” And it is clear that what was of concern in Fuentes was the private use of state power to achieve a nonconsensual resolution of a commercial dispute. The state statutes placed the state power to repossess property in the hands of an interested private party, just as the state statute in this case places the state power to conduct judicially binding sales in satisfaction of a lien in the hands of the warehouseman. . . .

This same point was made, equally emphatically, in Mitchell and Di-Chem. Yet the very defect that made the statutes in Fuentes and Di-Chem unconstitutional—lack of state control—is, under today’s decision, the factor that precludes constitutional review of the state statute. The Due Process Clause cannot command such incongruous results. If it is unconstitutional for a State to allow a private party to exercise a traditional state power because the state supervision of that power is purely mechanical, the State surely cannot immunize its actions from constitutional scrutiny by removing even the mechanical supervision.

Not only has the State removed its nominal supervision in this case,13 it has also authorized a private party to exercise a governmental power that is equally, if not more, significant than the power exercised in Fuentes or Di-Chem. In Fuentes, the Florida statute allowed the debtor’s property to be seized and held pending the outcome of the creditor’s action for repossession. The property would not be finally disposed of until there was an adjudication of the underlying claim. Similarly, in Di-Chem, the state statute provided for a garnishment procedure which deprived the debtor of the use of property in the garnishee’s hands pending the outcome of litigation. The warehouseman’s power under § 7-210 is far broader . . . .

13 Of course, the State does “supervise” the warehouseman’s actions in the sense that it prescribes the procedures that warehousemen must follow to complete a legally binding sale.
Whether termed “traditional,” “exclusive,” or “significant,” the state power to order binding, nonconsensual resolution of a conflict between debtor and creditor is exactly the sort of power with which the Due Process Clause is concerned. And the State’s delegation of that power to a private party is, accordingly, subject to due process scrutiny. This, at the very least, is the teaching of *Sniadach, Fuentes,* and *Di-Chem.*

It is important to emphasize that, contrary to the Court’s apparent fears, this conclusion does not even remotely suggest that “all private deprivations of property [will] be converted into public acts whenever the State, for whatever reason, denies relief sought by the putative property owner.” The focus is not on the private deprivation but on the state authorization. “[W]hat is always vital to remember is that it is the state’s conduct, whether action or inaction, not the private conduct, that gives rise to constitutional attack.” The State’s conduct in this case takes the concrete form of a statutory enactment, and it is that statute that may be challenged.

My analysis in this case thus assumes that petitioner Flagg Brothers’ proposed sale will conform to the procedure specified by the state legislature and that respondents’ challenge therefore will be to the constitutionality of that process. It is only what the State itself has enacted that they may ask the federal court to review in a § 1983 case. If there should be a deviation from the state statute—such as a failure to give the notice required by the state law—the defect could be remedied by a state court and there would be no occasion for § 1983 relief. [See *The Civil Rights Cases.*] 14

On the other hand, if there is compliance with the New York statute, the state legislative action which enabled the deprivation to take place must be subject to constitutional challenge in a federal court. 15 Under this approach, the federal courts do not have jurisdiction to review every foreclosure proceeding in which the debtor claims that there has been a procedural defect constituting a denial of due process of law. Rather, the federal district court’s jurisdiction under § 1983 is limited to challenges to the constitutionality of the state procedure itself—challenges of the kind considered in *Di-Chem* and *Fuentes.*

Finally, it is obviously true that the overwhelming majority of disputes in our society are resolved in the private sphere. But it is no longer possible, if it ever was, to believe that a sharp line can be drawn between private and public actions. The Court today holds that our examination of state delegations of power should be limited to those rare instances where the State has ceded one of its “exclusive” powers. As indicated, I believe that this limitation is neither logical nor practical. More troubling, this description of what is state action does not even attempt to reflect the concerns of the Due Process Clause, for the state-action doctrine is, after all, merely one aspect of this broad constitutional protection.

---

14 Furthermore, if the warehouseman has deviated from the statutory requirements, the statute would not provide him with the kind of support that would justify the conclusion that he acted “under color of law.” With respect to this requirement of § 1983, while I agree with the majority that the concepts of “under color of law” and “state action” may be separately analyzed, normally as a practical matter they embody the same test of state involvement.

15 Indeed, under the Court’s analysis as I understand it, the state statute in this case would not be subject to due process scrutiny in a state court.
In the broadest sense, we expect government “to provide a reasonable and fair framework of rules which facilitate commercial transactions . . . .” This “framework of rules” is premised on the assumption that the State will control nonconsensual deprivations of property and that the State’s control will, in turn, be subject to the restrictions of the Due Process Clause. The power to order legally binding surrenders of property and the constitutional restrictions on that power are necessary corollaries in our system. In effect, today’s decision allows the State to divorce these two elements by the simple expedient of transferring the implementation of its policy to private parties. Because the Fourteenth Amendment does not countenance such a division of power and responsibility, I respectfully dissent.


JUSTICE WHITE delivered the opinion of the Court.

I

. . . [P]etitioner, a lessee-operator of a truckstop . . . , was indebted to his supplier, Edmondson Oil Co., Inc. Edmondson sued on the debt in . . . state court [and] sought prejudgment attachment of certain of petitioner’s property. The prejudgment attachment procedure required only that Edmondson allege, in an ex parte petition, a belief that petitioner . . . might dispose of his property in order to defeat his creditors. Acting upon that petition, a Clerk of the state court issued a writ of attachment, which was then executed by the County Sheriff. This effectively sequestered petitioner’s property . . . . Pursuant to the statute, a hearing on the propriety of the attachment and levy was later conducted. Thirty-four days after the levy, a state trial judge ordered the attachment dismissed because Edmondson had failed to establish the statutory grounds for attachment alleged in the petition.

Petitioner subsequently brought this action under 42 U.S.C. § 1983 against Edmondson and its president[, seeking damages]. His complaint alleged that in attaching his property respondents had acted jointly with the State to deprive him of his property without due process of law. . . .

II

A

. . . In United States v. Price (1966), we explicitly stated that the requirements [of “state action” for the Fourteenth Amendment and “under color of state law” for § 1983] were identical . . . . In support of this proposition the Court cited Smith v. Allwright (1944) and Terry v. Adams (1953). In both of these cases black voters in Texas challenged their exclusion from party primaries as a violation of the Fifteenth Amendment and sought relief under [§ 1983]. In each case, the Court understood the problem before it to be whether the discriminatory policy of a private political association could be characterized

Draft — do not distribute!
as “state action . . . .” Having found state action . . . , there was no further inquiry into whether the action of the political associations also met the statutory requirement of action “under color of state law.”

Similarly, it is clear that in a § 1983 action brought against a state official, the statutory requirement of action “under color of state law” and the “state action” requirement of the Fourteenth Amendment are identical. . . .

. . . In Flagg Bros., the Court distinguished two elements of a § 1983 action:

“[Plaintiffs] are first bound to show that they have been deprived of a right ‘secured by the Constitution and the laws’ of the United States. They must secondly show that Flagg Brothers deprived them of this right acting ‘under color of any [state] statute’ . . . . It is clear that these two elements denote two separate areas of inquiry.”

. . . Th[is] two-part approach . . . was derived from Adickes v. S.H. Kress & Co. (1970). Adickes was a § 1983 action brought against a private party, based on a claim of racial discrimination in violation of the Equal Protection Clause of the Fourteenth Amendment. . . . [W]e held that the private party’s joint participation with a state official in a conspiracy to discriminate would constitute both “state action . . .” and action “‘under color’ of law . . . .” . . .

B

. . . Beginning with Sniadach, the Court has consistently held that constitutional requirements of due process apply to garnishment and prejudgment attachment procedures whenever officers of the State act jointly with a creditor in securing the property in dispute. Sniadach and Di-Chem involved state-created garnishment procedures; Mitchell involved execution of a vendor’s lien to secure disputed property. In each of these cases state agents aided the creditor in securing the disputed property; but in each case the federal issue arose in litigation between creditor and debtor in the state courts and no state official was named as a party. Nevertheless, in each case the Court entertained and adjudicated the defendant-debtor’s claim that the procedure under which the private creditor secured the disputed property violated . . . due process. Necessary to that conclusion is the holding that private use of the challenged state procedures with the help of state officials constitutes state action . . . .

Fuentes was a § 1983 action brought against both a private creditor and the State Attorney General. The plaintiff sought declaratory and injunctive relief, on due process grounds, from continued enforcement of state statutes authorizing prejudgment replevin. The plaintiff prevailed; . . . the private parties . . . remained parties, and judgment ran against them in this Court. 16

16 We thus find incomprehensible JUSTICE POWELL’s statement that we cite no cases in which a private decision to invoke a presumptively valid state legal process has been held to be state action. Likewise, his discussion of these cases steadfastly ignores the predicate for the holding in each case that the debtor could
If a defendant debtor in state-court debt collection proceedings can successfully challenge, on federal due process grounds, the plaintiff creditor’s resort to the procedures authorized by a state statute, it is difficult to understand why that same behavior by the state-court plaintiff should not provide a cause of action under § 1983. If the creditor-plaintiff violates the debtor-defendant’s due process rights by seizing his property in accordance with statutory procedures, there is little or no reason to deny to the latter a cause of action under the federal statute, § 1983, designed to provide judicial redress for just such constitutional violations.

To read the “under color of any statute” language of the Act in such a way as to impose a limit on those Fourteenth Amendment violations that may be redressed by the § 1983 cause of action would be wholly inconsistent with the purpose of § 1 of the Civil Rights Act of 1871, from which § 1983 is derived. The Act was passed “for the express purpose of ‘enforc[ing] the Provisions of the Fourteenth Amendment.’” The history of the Act is replete with statements indicating that Congress thought it was creating a remedy as broad as the protection that the Fourteenth Amendment affords the individual. . . .

In sum, . . . [i]f the challenged conduct of respondents constitutes state action as delimited by our prior decisions, then that conduct was also action under color of state law and will support a suit under § 1983.

III

As a matter of substantive constitutional law the state-action requirement reflects . . . the fact that “most rights secured by the Constitution are protected only against infringement by governments.” . . .

Our cases have accordingly insisted that the conduct allegedly causing the deprivation of a federal right be fairly attributable to the State. These cases reflect a two-part approach to this question of “fair attribution.” First, the deprivation must be caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the state or by a person for whom the State is responsible. In Sniadach, Fuentes, W.T. Grant, and North Georgia, for example, a state statute provided the right to garnish or to obtain prejudgment attachment, as well as the procedure by which the rights could be exercised. Second, the party charged with the deprivation must be a person who may fairly be said to be a state actor. This may be because he is a state official, because he has acted together with or has obtained significant aid from state officials, or because his conduct is otherwise chargeable to the State. Without a limit such as this, private parties could face constitutional litigation whenever they seek to rely on some state rule governing their interactions with the community surrounding them.

Although related, these two principles are not the same. They collapse into each other when the claim of a constitutional deprivation is directed against a party whose official challenge the constitutional adequacy of the private creditor’s seizure of his property. That predicate was necessarily the principle that a private party’s invocation of a seemingly valid prejudgment remedy statute, coupled with the aid of a state official, satisfies the state-action requirement of the Fourteenth Amendment and warrants relief against the private party.
character is such as to lend the weight of the State to his decisions. The two principles diverge when the constitutional claim is directed against a party without such apparent authority, *i.e.*, against a private party. The difference between the two inquiries is well illustrated by comparing *Moose Lodge No. 107 v. Irvis* (1972), with *Flagg Brothers*.

In *Moose Lodge*, the Court held that the discriminatory practices of the appellant did not violate the Equal Protection Clause because those practices did not constitute “state action.” The Court focused primarily on the question of whether the admittedly discriminatory policy could in any way be ascribed to a governmental decision. The inquiry, therefore, looked to those policies adopted by the State that were applied to appellant. The Court concluded as follows:

“We therefore hold, that with the exception hereafter noted, the operation of the regulatory scheme enforced by the Pennsylvania Liquor Control Board does not sufficiently implicate the State in the discriminatory guest policies of Moose Lodge to . . . make the latter ‘state action’ . . . .”

In other words, the decision to discriminate could not be ascribed to any governmental decision; those governmental decisions that did affect Moose Lodge were unconnected with its discriminatory policies.\(^{20}\)

*Flagg Brothers* focused on the other component of the state-action principle. In that case, the warehouseman proceeded under [a state statute], and the debtor challenged the constitutionality of that provision [under] the Due Process and Equal Protection Clauses . . . . Undoubtedly the State was responsible for the statute. The response of the Court, however, focused not on the terms of the statute but on the character of the defendant to the § 1983 suit: Action by a private party pursuant to this statute, without something more, was not sufficient to justify a characterization of that party as a “state actor.” The Court suggested that that “something more” which would convert the private party into a state actor might vary with the circumstances of the case. This was simply a recognition that the Court has articulated a number of different factors or tests in different contexts: *e.g.*, the “public function” test, the “state compulsion” test, the “nexus” test, and in the case of prejudgment attachments, a “joint action test.”\(^{21}\) Whether these different tests are actually different in operation or simply different ways of characterizing the necessarily fact-bound inquiry that confronts the Court in such a situation need not be resolved here.

IV

Turning to this case, the first question is whether the claimed deprivation has resulted from the exercise of a right or privilege having its source in state authority. The second

---

\(^{20}\) The “one exception” further illustrates this point. The Court enjoined enforcement of a state rule requiring Moose Lodge to comply with its own constitution and bylaws insofar as they contained racially discriminatory provisions. State enforcement of this rule, either judicially or administratively, would, under the circumstances, amount to a governmental decision to adopt a racially discriminatory policy.

\(^{21}\) Contrary to the suggestion of JUSTICE POWELL’s dissent, we do not hold today that “a private party’s mere invocation of state legal procedures constitutes ‘joint participation’ or ‘conspiracy’ with state officials satisfying the § 1983 requirement of action under color of law.” The holding today, as the above analysis makes clear, is limited to the particular context of prejudgment attachment.
question is whether, under the facts of this case, respondents, who are private parties, may be appropriately characterized as “state actors.”

... [We read his complaint as challenging] the state statute as procedurally defective under the Fourteenth Amendment.

... [T]he procedural scheme created by the statute obviously is the product of state action. This is subject to constitutional restraints and properly may be addressed in a § 1983 action, if the second element of the state-action requirement is met as well.

As is clear from the discussion in Part II, we have consistently held that a private party’s joint participation with state officials in the seizure of disputed property is sufficient to characterize that party as a “state actor” for purposes of the Fourteenth Amendment. The rule in these cases is the same as that articulated... in the context of an equal protection deprivation:

“Private persons, jointly engaged with state officials in the prohibited action, are acting “under color” of law for purposes of the statute. To act “under color” of law does not require that the accused be an officer of the State. It is enough that he is a willful participant in joint activity with the State or its agents.”

The Court of Appeals erred in holding that in this context “joint participation” required something more than invoking the aid of state officials to take advantage of state-created attachment procedures. Whatever may be true in other contexts, this is sufficient when the State has created a system whereby state officials will attach property on the ex parte application of one party to a private dispute.

In summary, petitioner was deprived of his property through state action; respondents were, therefore, acting under color of state law in participating in that deprivation. ... 23

...  

CHIEF JUSTICE BURGER, dissenting.

Whether we are dealing with suits under § 1983 or suits brought pursuant to the Fourteenth Amendment, in my view the inquiry is the same: is the claimed infringement of a federal right fairly attributable to the State. Applying this standard, it cannot be said that the actions of the named respondents are fairly attributable to the State.” Respondents did no more than invoke a presumptively valid state prejudgment attachment procedure...

23 JUSTICE POWELL is concerned that private individuals who innocently make use of seemingly valid state laws would be responsible, if the law is subsequently held to be unconstitutional, for the consequences of their actions. In our view, however, this problem should be dealt with not by changing the character of the cause of action but by establishing an affirmative defense. A similar concern is at least partially responsible for the availability of a good-faith defense, or qualified immunity, to state officials. We need not reach the question of the availability of such a defense to private individuals at this juncture. ...
available to all. Relying on a dubious “but for” analysis, the Court erroneously concludes that the subsequent procedural steps taken by the State in attaching a putative debtor’s property in some way transforms respondents’ acts into actions of the State. This case is no different from the situation in which a private party commences a lawsuit and secures injunctive relief which, even if temporary, may cause significant injury to the defendant. Invoking a judicial process, of course, implicates the State and its officers but does not transform essentially private conduct into actions of the State. Similarly, one who practices a trade or profession, drives an automobile, or builds a house under a state license is not engaging in acts fairly attributable to the state. . . . [P]etitioner’s remedy lies in private suits for damages such as malicious prosecution. The Court’s opinion expands the reach of the statute beyond anything intended by Congress. It may well be a consequence of too casually falling into a semantical trap because of the figurative use of the term “color of state law.”

JUSTICE POWELL, with whom JUSTICE REHNQUIST and JUSTICE O’CONNOR join, dissenting.

Today’s decision is a disquieting example of how expansive judicial decisionmaking can ensnare a person who had every reason to believe he was acting in strict accordance with law. . . . No court had questioned the validity of the statute [which goes back to 1819] . . . . There is no allegation that respondent conspired with the state officials to deny petitioner the fair protection of state or federal law.

Respondent ultimately prevailed in his lawsuit. The petitioner Lugar was ordered by a court to pay his debt. A state court did find, however, that Lugar’s assets should not have been attached prior to a judgment on the underlying action. . . .

This Court today . . . holds that respondent, a private citizen who did no more than commence a legal action of a kind traditionally initiated by private parties, thereby engaged in “state action.” This decision is as unprecedented as it is implausible. It is plainly unjust to the respondent, and the Court makes no argument to the contrary. Respondent, who was represented by counsel, could have had no notion that his filing of a petition in state court, in the effort to secure payment of a private debt, made him a “state actor” liable in damages for allegedly unconstitutional action by the Commonwealth of Virginia. . . . [T]he Court undermines fundamental distinctions between the common-sense categories of state and private conduct and between the legal concepts of “state action” and private action “under color of law.”

I

The plain language of 42 U.S.C. § 1983 establishes that a plaintiff must satisfy two jurisdictional requisites to state an actionable claim. First, he must allege the violation of a right “secured by the Constitution and laws” of the United States. Because “most rights secured by the Constitution are protected only against infringement by governments,” this requirement compels an inquiry into the presence of state action. Second, a § 1983 plaintiff must show that the alleged deprivation was caused by a person acting “under
color” of law. In Flagg Bros., this Court affirmed that “these two elements denote two separate areas of inquiry.”

This case demonstrates why separate inquiries are required. Here it is not disputed that the Virginia Sheriff and Clerk of Court, the state officials who sequestered petitioner’s property in the manner provided by Virginia law, engaged in state action. Yet the petitioner, while alleging constitutional injury from this action by state officials, did not sue the State or its agents. In these circumstances the Court of Appeals correctly stated that the relevant inquiry was the second identified in Flagg Bros.: whether the respondent, a private citizen whose only action was to invoke a presumptively valid state attachment process, had acted under color of state law in “causing” the State to deprive petitioner of alleged constitutional rights. . . .

. . . [T]he Court opinion inexplicably conflates the[se] two inquiries . . . . Ignoring that this case involves two sets of actions—one by respondent, who merely filed a suit and accompanying sequestration petition; another by the state officials, who issued the writ and executed the lien—it wrongly frames the question before the Court, not as whether the private respondent acted under color of law in filing the petition, but as “whether . . . respondents, who are private parties, may be appropriately characterized as ‘state actors.’” It then concludes that they may, on the theory that a private party who invokes “the aid of state officials to take advantage of state-created attachment procedures” is a “joint participant” with the State and therefore a “state actor.” . . .

There are at least two fallacies in the Court’s conclusion. First, . . . our cases have not established that private “joint participants” with state officials themselves necessarily become state actors. Where private citizens interact with state officials in the pursuit of merely private ends, the appropriate inquiry generally is whether the private parties have acted “under color of law.” Second, even when the inquiry is whether an action occurred under color of law, our cases make clear that the “joint participation” standard is not satisfied when a private citizen does no more than invoke a presumptively valid judicial process in pursuit only of legitimate private ends.

II

[T]he historic purpose of § 1983 was to prevent state officials from using the cloak of their authority under state law to violate rights protected against state infringement by the Fourteenth Amendment. The Court accordingly is correct that an important inquiry in a § 1983 suit against a private party is whether there is an allegation of wrongful “conduct that can be attributed to the State.” This is the first question referred to in Flagg Bros. But there still remains the second Flagg Bros. question: whether this state action fairly can be attributed to the respondent, whose only action was to invoke a presumptively valid attachment statute. This question, unasked by the Court, reveals the fallacy of its conclusion that respondent may be held accountable for the attachment of property because he was a “state actor.” From the occurrence of state action taken by the Sheriff who sequestered petitioner’s property, it does not follow that respondent became a “state actor” simply because the Sheriff was. This Court, until today, has never endorsed this non sequitur.
It of course is true that respondent’s private action was followed by state action, and that the private and the state actions were not unconnected. But “[t]hat the State responds to [private] actions by [taking action of its own] does not render it responsible for those [private] actions.” And where the State is not responsible for a private decision to behave in a certain way, the private action generally cannot be considered “state action” within the meaning of our cases. . . .

This Court of course has held that private parties are amenable to suit under § 1983 when “jointly engaged” with state officials in the violation of constitutional rights. Yet the Court, in advancing its “joint participation” theory, does not cite a single case in which a private decision to invoke a presumptively valid state legal process has been held to constitute state action. Even the quotation on which the Court principally relies for its statement of the applicable “rule” does not refer to state action. Rather, it states explicitly that “[p]rivate persons, jointly engaged with state officials in the prohibited action, are acting ‘under color’ of law for purposes of the statute.”

. . . [The] distinction between “state action” and private action under “color of law” . . . is sound in principle. It also is consistent with and supportive of the distinction between “private” conduct and government action that is subject to the procedural limitations of the Due Process Clause . . . . “Careful adherence to the ‘state action’ requirement preserves an area of individual freedom by limiting the reach of federal law and federal judicial power. It also avoids imposing on the State, its agencies or officials, responsibility for conduct for which they cannot fairly be blamed.”

A “color of law” inquiry acknowledges that private individuals, engaged in unlawful joint behavior with state officials, may be personally responsible for wrongs that they cause to occur. But it does not confuse private actors with the State—the fallacy of the analysis adopted today by the Court. In this case involving the private action of the respondent in petitioning the state courts . . . , the appropriate inquiry as to respondent’s liability is not whether he was a state actor, but whether he acted under color of law. It is to this question that I therefore turn.

III

Contrary to the position of the Court, our cases do not establish that a private party’s mere invocation of state legal procedures constitutes “joint participation” or “conspiracy” with state officials satisfying the § 1983 requirement of action under color of law. . . . [I]nvocation of state legal process is “essentially a private function . . . for which state office and authority are not needed.” . . . [I]ndependent, private decisions made in the context of litigation cannot be said to occur under color of law. The Court nevertheless advances two principal grounds for its holding to the contrary.

---

7 In *Adickes* the term “jointly engaged” appears to have been used specifically to connote engagement in a “conspiracy.”
A

The Court argues that petitioner’s action under § 1983 is supported by cases in which this Court has applied due process standards to state garnishment and prejudgment attachment procedures. The Court relies specifically on Sniadach, Fuentes v. Shevin (1972), Mitchell, and Di-Chem. According to the Court, these cases establish that a private party acts “under color” of law when seeking the attachment of property under an unconstitutional state statute. In fact, a careful reading demonstrates that they provide no authority for this proposition.

Of the cases cited by the Court, Sniadach, Mitchell, and Di-Chem all involved attacks on the validity of state attachment or garnishment statutes. None of the cases alleged that the private creditor was a joint actor with the State, and none involved a claim for damages against the creditor. Each case involved a state suit, not a federal action under § 1983. It therefore was unnecessary in any of these cases for this Court to consider whether the creditor, by virtue of instituting the attachment or garnishment, became a state actor or acted under color of state law. There is not one word in any of these cases that so characterizes the private creditor. In Fuentes, the Court did consider a § 1983 action against a private creditor as well as the State Attorney General. Again, however, the only question before this Court was the validity of a state statute. No claim was made that the creditor was a joint actor with the State or had acted under color of law. No damages were sought from the creditor. Again, there was no occasion for this Court to consider the status under § 1983 of the private party, and there is not a word in the opinion that discusses this. . . .

B

In addition to relying on cases involving the constitutionality of state attachment and garnishment statutes, the Court advances a “joint participation” theory based on Adickes. In Adickes the plaintiff sued a private restaurant under § 1983, alleging a conspiracy between the restaurant and local police to deprive her of the right to equal treatment in a place of public accommodation. . . . [T]his Court . . . found that the private defendant, in “conspiring” with local police to obtain official enforcement of a state custom of racial segregation, engaged in a “‘joint activity with the State or its agents’” and therefore acted under color of law within the meaning of § 1983.

Contrary to the suggestion of the Court, however, . . . Adickes did not purport to define the term “under color of law.” Attending closely to the facts presented, the Court observed that “[w]hatever else may also be necessary to show that a person has acted ‘under color of [a] statute’ for purposes of § 1983, . . . we think it essential that he act with the knowledge of and pursuant to that statute.” As indicated by this choice of language, the Court clearly seems to have contemplated some limiting principle. A citizen summoning the police to enforce the law ordinarily would not be considered to have engaged in a “conspiracy.” Nor, presumably, would such a citizen be characterized as acting under color of law and thereby risking amenability to suit for constitutional violations that subsequently might occur. Surely there is nothing in Adickes to indicate that the Court would have found action under color of law in cases of this kind.
. . . Adickes [is best read as] establish[ing] that a private party acts under color of law when he conspires with state officials to secure the application of a state law so plainly unconstitutional as to enjoy no presumption of validity. . . . Here, however, petitioner has alleged no conspiracy. Nor has he even alleged that respondent was invoking the aid of a law he should have known to be constitutionally invalid. Finally, there is no allegation that respondent’s decision to invoke legal process was in any way compelled by the law or custom of the State in which he lived. In this context Adickes simply is inapposite. . . .

§ 2.6. Notes on the “individual remedies” cases

1. Adickes v. S.H. Kress & Co. (1970). Adickes, a white teacher at a Mississippi “Freedom School,” came to a lunch counter at the Kress store with six black students. The students were served, but she was not because she was “a white person ‘in the company of Negroes.’” When she left the store, she was arrested for vagrancy by a policeman who had observed the incident at the lunch counter. She sued the lunch counter, arguing that the refusal to serve her was pursuant to a “custom of the community to segregate the races in public eating places.”

The Court stated that “our cases make clear that” the Fourteenth Amendment is violated if “a Kress employee, in the course of employment, and a . . . policeman somehow reached an understanding to deny Miss Adickes service in the Kress store, or to cause her subsequent arrest because she was a white person in the company of Negroes. The involvement of a state official in such a conspiracy plainly provides the state action essential to show a direct violation of” Equal Protection. “For state action purposes it makes no difference whether the racially discriminatory act by the private party is compelled by a statutory provision or by a custom having the force of law—in either case it is the State that has commanded the result by its law.” See also the “Government compulsion or influence” cases in note 2 of § 2.10.

2. The involvement of state officials. In Justice Rehnquist’s view, Flagg Bros. is distinguishable from prior cases because no state officials were involved. Would the outcome have been different if the parties had not agreed to dismiss the City Marshal who supervised the evictions? What if the statute didn’t provide for supervision by the City Marshal—would there be no way to challenge the statute? Behind every authorized act, there stands the implicit promise of judicial enforcement; why should the absence of overt official involvement be important?

14 The Court suggests that respondent may be entitled to claim good-faith immunity from this suit for civil damages. This is a positive suggestion with which I agree. A holding of immunity will mitigate the ultimate cost of this litigation. It would not, however, convert the Court’s holding into a just one. This case already has been in litigation for nearly five years. It will now be remanded for further proceedings. Respondent, solely because he undertook to assert rights authorized by a presumptively valid state statute, will have been subjected to the expense, distractions, and hazards of a protracted litigation.
3. **Consent vs. coercion.** It may be tempting to justify the outcome in *Flagg Bros.* on the grounds that the New York statute only set a waivable contractual default—thus, the remedy at issue became part of the private contract and was therefore not state action. However, the case was not decided on this ground. As Justice Stevens points out, the lower court explicitly rejected any contractual foundation for the case, and certainly Brooks’s co-petitioner Jones claimed that she never agreed to any storage at all. We must assume, for the purpose of this case, that the sale was not authorized by the statute but not agreed to.

Rehnquist suggests (n.10) that the New York statute merely redefines property interests so that the warehouseman has a property right in the items he was selling—and the mere exercise of a property isn’t state action. Stevens responds (n.3 & following ¶ in text) that this theory seems to have no attractive stopping point. Rehnquist seeks to avoid this problem (n.9) by limiting the holding to commercial dispute resolution, but does this limitation make sense?

4. **Exclusivity.** Later cases discuss the “exclusive public function” test—for instance, education of problem children (*Rendell-Baker*) or operation of nursing homes (*Blum*) are not exclusive public functions, while imprisonment (*West*, as characterized in *Sullivan*) is. Here, Rehnquist finds non-exclusivity because there are other remedies available to petitioners. Stevens finds this “incomprehensible” (n.8). Do you find his critique convincing?

5. **The need to pin the blame.** When someone challenges an action as being unconstitutional, *whether* there was state action is often more important than *who* was the state actor. In cases like *Sniadach, Mitchell*, and *Di-Chem*, as the Powell dissent explains, state officials were involved in the enforcement of various creditors’ remedies, so there was enough state action to be able to subject the remedies to Due Process analysis—but because all that was sought was the non-enforcement of the remedy, it was not necessary to determine whether the state actors were the state officials only, or also the creditors.

In cases seeking damages under 42 U.S.C. § 1983, the identity question becomes much more important, because one can only collect damages from someone acting “under color of law,” and only if they subject someone (or cause someone to be subjected) to a deprivation of constitutional rights. In *Flagg Bros.*, the majority found no state action on the part of anyone, and therefore didn’t reach the “under color of law” inquiry; thus, as the dissent argues (n.15), under the majority’s analysis, the law is entirely immune from Due Process scrutiny. The dissent, on the other hand, found state action on behalf of both the warehouseman and the legislature (and suggested that the “under color of law” inquiry “normally as a practical matter . . . embod[ies] the same test”). Thus, *Flagg Bros.* becomes subject not only to Due Process restrictions but also to Equal Protection and First Amendment restrictions—one could in principle require it to sue blacks and whites, Republicans and Democrats, evenhandedly. Is this a desirable result? In *Lugar*, the majority finds that the private party is both a state actor and acts under color of state law; the dissent would have found the opposite on each count.
Couldn’t one disentangle the two inquiries even more? The judicial enforcers (and possibly legislative enacters) would be the state actors in every instance—because behind any authorized action lies the implicit promise of judicial enforcement. As for the private parties, they would never be state actors, but they could cause someone to be subjected to a deprivation of constitutional rights by the state actors. Moreover, to the extent they act under the authorization of a statute, they could be said to be acting “under color of state law.” This theory would make the private parties acceptable § 1983 plaintiffs in both Flagg Bros. and Lugar, but by not making them state actors, it would leave them free to choose whether to invoke state process on whatever grounds they chose, without having to worry about being subject to Equal Protection or First Amendment restrictions.

6. The injustice of Lugar. Justice Powell makes the injustice of holding a litigant responsible for the constitutionality of a well-entrenched legal procedure a theme of his Lugar dissent. Justice White argues that the response to this concern would be to create a good-faith defense (n.23); Powell agrees (n.14), but suggests this is not good enough because the key is to avoid litigation in the first place, not to win after one has been dragged through years of litigation. Why would one have to wait years to benefit from such a defense; couldn’t it be applied early in the case? For more on this, see the discussion of immunity doctrines in a later section of these materials.


Justice Rehnquist delivered the opinion of the Court.

Respondents represent a class of Medicaid patients challenging decisions by the nursing homes in which they reside to discharge or transfer patients without notice or an opportunity for a hearing. The question is whether the State may be held responsible for those decisions so as to subject them to the strictures of the Fourteenth Amendment.

I

Congress established the Medicaid program in 1965 . . . to provide federal financial assistance to States that choose to reimburse certain medical costs incurred by the poor. As a participating State, New York provides Medicaid assistance to eligible persons who receive care in private nursing homes, which are designated as either “skilled nursing facilities” (SNF’s) or “health related facilities” (HRF’s). The latter provide less extensive, and generally less expensive, medical care than the former. Nursing homes chosen by Medicaid patients are directly reimbursed by the State for the reasonable cost of health care services.

An individual must meet two conditions to obtain Medicaid assistance. He must satisfy eligibility standards defined in terms of income or resources and he must seek medically
necessary services. To assure that the latter condition is satisfied, federal regulations require each nursing home to establish a utilization review committee (URC) of physicians whose functions include periodically assessing whether each patient is receiving the appropriate level of care, and thus whether the patient’s continued stay in the facility is justified. If the URC determines that the patient should be discharged or transferred to a different level of care, either more or less intensive, it must notify the state agency responsible for administering Medicaid assistance.

. . . Yaretsky . . . [was a Medicaid beneficiary]. . . . [His] nursing home’s URC decided that [he] did not need the care [he was] receiving and should be transferred to a lower level of care in an HRF. New York City officials, who were then responsible for administering the Medicaid program in the city, were notified of this decision and prepared to reduce or terminate payments to the nursing home for respondents’ care. Following administrative hearings, state social service officials affirmed the decision to discontinue benefits unless respondents accepted a transfer to an HRF providing a reduced level of care.

[Yaretsky sued] . . . the Commissioners of the New York Department of Social Services and the Department of Health[, alleging due process violations]. . . .

III

. . . “[T]he action inhibited by the first section of the Fourteenth Amendment is only such action as may fairly be said to be that of the States. . . . That Amendment erects no shield against merely private conduct, however discriminatory or wrongful.” Shelley.

Faithful adherence to the “state action” requirement of the Fourteenth Amendment requires careful attention to the gravamen of the plaintiff’s complaint. In this case, respondents objected to the involuntary discharge or transfer of Medicaid patients by their nursing homes without certain procedural safeguards. They have named as defendants state officials responsible for administering the Medicaid program in New York. These officials are also responsible for regulating nursing homes in the State, including those in which respondents were receiving care. But respondents are not challenging particular state regulations or procedures, and their arguments concede that the decision to discharge or transfer a patient originates not with state officials, but with nursing homes that are privately owned and operated. Their lawsuit, therefore, seeks to hold state officials liable for the actions of private parties, and the injunctive relief they have obtained requires the State to adopt regulations that will prohibit the private conduct of which they complain.

A

This case is obviously different from those cases in which the defendant is a private party and the question is whether his conduct has sufficiently received the imprimatur of the

4 These committees must be composed of private physicians who are not directly responsible for the patient whose care is being reviewed. Under New York law, the committee members may not be employed by the SNF or HRF and may not have a financial interest in any residential care facility.

125

Draft — do not distribute!
State so as to make it “state” action for purposes of the Fourteenth Amendment. It also differs from other “state action” cases in which the challenged conduct consists of enforcement of state laws or regulations by state officials who are themselves parties in the lawsuit; in such cases the question typically is whether the private motives which triggered the enforcement of those laws can fairly be attributed to the State. But both these types of cases shed light upon the analysis necessary to resolve the present case.

First, although it is apparent that nursing homes in New York are extensively regulated, “[t]he mere fact that a business is subject to state regulation does not by itself convert its action into that of the State for purposes of the Fourteenth Amendment.” The complaining party must also show that “there is a sufficiently close nexus between the State and the challenged action of the regulated entity so that the action of the latter may be fairly treated as that of the State itself.” The purpose of this requirement is to assure that constitutional standards are invoked only when it can be said that the State is responsible for the specific conduct of which the plaintiff complains. The importance of this assurance is evident when, as in this case, the complaining party seeks to hold the State liable for the actions of private parties.

Second, although the factual setting of each case will be significant, our precedents indicate that a State normally can be held responsible for a private decision only when it has exercised coercive power or has provided such significant encouragement, either overt or covert, that the choice must in law be deemed to be that of the State. Mere approval of or acquiescence in the initiatives of a private party is not sufficient to justify holding the State responsible for those initiatives under the terms of the Fourteenth Amendment.

Third, the required nexus may be present if the private entity has exercised powers that are “traditionally the exclusive prerogative of the State.”

B

. . . [The state action present was the adjustment of the patient’s benefits in response to the nursing homes’ discharge or transfer decisions.] . . . Respondents, however, do not challenge the adjustment of benefits, but the discharge or transfer of patients to lower levels of care without adequate notice or hearings. That the State responds to such actions by adjusting benefits does not render it responsible for those actions. The decisions about which respondents complain are made by physicians and nursing home administrators, all of whom are concededly private parties. There is no suggestion that those decisions were influenced in any degree by the State’s obligation to adjust benefits in conformity with changes in the cost of medically necessary care.

Respondents . . . [also] argue that the State “affirmatively commands” the summary discharge or transfer of Medicaid patients who are thought to be inappropriately placed in their nursing facilities. Were this characterization accurate, we would have a different question before us. However, our review of the statutes and regulations identified by respondents does not support respondents’ characterization of them.
As our earlier summary of the Medicaid program explained, a patient must meet two essential conditions in order to obtain financial assistance. He must satisfy eligibility criteria defined in terms of income and resources and he must seek medically necessary services. To assure that nursing home services are medically necessary, federal law requires that a physician certify at the time the Medicaid patient is admitted and periodically thereafter. New York requires that the physician complete a “long term care placement form” devised by the Department of Health, called the DMS-1. A completed form provides, inter alia, a numerical score corresponding to the physician’s assessment of the patient’s mental and physical health. As petitioners note, however, the physicians, and not the forms, make the decision about whether the patient’s care is medically necessary. A physician can authorize a patient’s admission to a nursing facility despite a “low” score on the form. We cannot say that the State, by requiring completion of a form, is responsible for the physician’s decision.

In any case, respondents’ complaint is about nursing home decisions to discharge or transfer, not to admit, Medicaid patients. But we are not satisfied that the State is responsible for those decisions either. The regulations cited by respondents require

15 A completed DMS-1 form provides a summary of the patient’s medical condition. Five of the eleven questions devoted to this subject require the assignment of numerical values. A range of numerical values to be used in completing these questions are set forth in a second form, called the DMS-9. The dissent’s discussion of the DMS-9 suggests that completion of the DMS-1 form is a purely mechanical exercise that does not require the exercise of independent medical judgment. The dissent’s discussion is incomplete. The other six questions on the DMS-1 ask the physician such questions as whether the patient requires daily supervision by a registered nurse, whether complications would arise without skilled nursing care, whether a program of therapy is necessary, and if so what kind, whether the patient should be considered for different levels of care, and whether the patient is medically qualified for the level of care he or she is receiving. The physician brings to bear his own medical judgment in answering these questions; their placement on the form would be inexplicable if the numerical scores were dispositive.

16 The dissent belittles this fact by noting that the decision to depart from the form in admitting a patient is made by a physician member of the nursing home’s URC, and that such persons are “part and parcel of the statutory cost control process.” This signifies nothing more than the fact, disputed by no one, that the State requires utilization review in order to reduce unnecessary Medicaid expenditures. It remains true that physician members of the URC’s are not employed by the State and, more important, render medical judgments concerning the patient’s health needs that the State does not prescribe and for which it is not responsible. We must also emphasize, of course, that we are ultimately concerned with decisions to transfer patients who have already been admitted. Apropos of this relevant issue, the dissent observes that once a patient has been admitted, the State requires, as a condition to the disbursement of Medicaid funds, that within five days after admission the nursing home operator assess the patient’s status according to standards contained in the DMS-1 and DMS-9 forms. As the dissent is also aware, a physician member of the URC has the power to determine that the patient needs the level of care he is receiving despite an adverse score on the DMS-1. That decision, rendered after consultation with the patient’s attending physician, is purely a medical judgment for which the State, as before, is not responsible.

17 The dissent condemns us for conducting a “cursory” review of the regulations governing utilization review and pointedly asks “where . . . is the Court's discussion of the frequent utilization reviews that occur after admission?” The dissent, in its headlong dive into the sea of state regulations, forgets that patient transfers to lower levels of care initiated by utilization review committees are simply not part of this case. Such transfers were the subject of a consent judgment in October 1979. We are concerned only with transfers initiated by the patients’ attending physicians or the nursing home administrators themselves. Therefore, we have focused on regulations that concern decisions which are not the product of URC recommendations. As we explain in the text, those regulations do not demonstrate that the State is responsible for the transfers with which we are concerned.

Draft — do not distribute!
SNF’s and HRF’s “to make all efforts possible to transfer patients to the appropriate level of care or home as indicated by the patient’s medical condition or needs.” The nursing homes are required to complete patient care assessment forms designed by the State and “provide the receiving facility or provider with a current copy of same at the time of discharge to an alternate level of care facility or home.”

These regulations do not require the nursing homes to rely on the forms in making discharge or transfer decisions, nor do they demonstrate that the State is responsible for the decision to discharge or transfer particular patients. Those decisions ultimately turn on medical judgments made by private parties according to professional standards that are not established by the State. 19 This case, therefore, is not unlike Polk County v. Dodson (1981), in which the question was whether a public defender acts “under color of” state law within the meaning of 42 U.S.C. § 1983 when representing an indigent defendant in a state criminal proceeding. Although the public defender was employed by the State and appointed by the State to represent the respondent, we concluded that “[t]his assignment entailed functions and obligations in no way dependent on state authority.” The decisions made by the public defender in the course of representing his client were framed in accordance with professional canons of ethics, rather than dictated by any rule of conduct imposed by the State. The same is true of nursing home decisions to discharge or transfer particular patients because the care they are receiving is medically inappropriate.

Respondents next point to regulations which, they say, impose a range of penalties on nursing homes that fail to discharge or transfer patients whose continued stay is inappropriate. One regulation excludes from participation in the Medicaid program health care providers who “[f]urnished items or services that are substantially in excess of the beneficiary’s needs.” The State is also authorized to fine health care providers who violate applicable regulations. As we have previously concluded, however, those regulations themselves do not dictate the decision to discharge or transfer in a particular case. Consequently, penalties imposed for violating the regulations add nothing to respondents’ claim of state action.

As an alternative position, respondents argue that even if the State does not command the transfers at issue, it reviews and either approves or rejects them on the merits. The regulations cited by respondents will not bear this construction. Although the State

19 The dissent characterizes as “factually unfounded” our conclusion that decisions initiated by nursing homes and physicians to transfer patients to lower levels of care ultimately depend on private judgments about the health needs of the patients. It asserts that different levels of care exist only because of the State’s desire to save money, and that the same interest explains the requirement that nursing homes transfer patients who do not need the care they are receiving. We do not suggest otherwise. Transfers to lower levels of care are not mandated by the patients’ health needs. But they occur only after an assessment of those needs. In other words, although “downward” transfers are made possible and encouraged for efficiency reasons, they can occur only after the decision is made that the patient does not need the care he or she is currently receiving. The State is simply not responsible for that decision, although it clearly responds to it. In concrete terms, therefore, if a particular patient objects to his transfer to a different nursing facility, the “fault” lies not with the State but ultimately with the judgment, made by concededly private parties, that he is receiving expensive care that he does not need. That judgment is a medical one, not a question of accounting.
requires the nursing homes to complete patient care assessment forms and file them with state Medicaid officials, and although federal law requires that state officials review these assessments, nothing in the regulations authorizes the officials to approve or disapprove decisions either to retain or discharge particular patients, and petitioners specifically disclaim any such responsibility. Instead, the State is obliged to approve or disapprove continued payment of Medicaid benefits after a change in the patient’s need for services. Adjustments in benefit levels in response to a decision to discharge or transfer a patient does not constitute approval or enforcement of that decision. As we have already concluded, this degree of involvement is too slim a basis on which to predicate a finding of state action in the decision itself.

Finally, respondents advance the rather vague generalization that such a relationship exists between the State and the nursing homes it regulates that the State may be considered a joint participant in the homes’ discharge and transfer of Medicaid patients. For this proposition they rely upon *Burton*. Respondents argue that state subsidization of the operating and capital costs of the facilities, payment of the medical expenses of more than 90% of the patients in the facilities, and the licensing of the facilities by the State, taken together convert the action of the homes into “state” action. But accepting all of these assertions as true, we are nonetheless unable to agree that the State is responsible for the decisions challenged by respondents. As we have previously held, privately owned enterprises providing services that the State would not necessarily provide, even though they are extensively regulated, do not fall within the ambit of *Burton*. That programs undertaken by the State result in substantial funding of the activities of a private entity is no more persuasive than the fact of regulation of such an entity in demonstrating that the State is responsible for decisions made by the entity in the course of its business.

We are also unable to conclude that the nursing homes perform a function that has been “traditionally the exclusive prerogative of the State.” Respondents’ argument in this regard is premised on their assertion that both the Medicaid statute and the New York Constitution make the State responsible for providing every Medicaid patient with nursing home services. The state constitutional provisions cited by respondents, however, do no more than authorize the legislature to provide funds for the care of the needy. They do not mandate the provision of any particular care, much less long-term nursing care. Similarly, the Medicaid statute requires that the States provide funding for skilled nursing services as a condition to the receipt of federal moneys. It does not require that the States provide the services themselves. Even if respondents’ characterization of the State’s duties were correct, however, it would not follow that decisions made in the day-to-day administration of a nursing home are the kind of decisions traditionally and exclusively made by the sovereign for and on behalf of the public. Indeed, respondents make no such claim, nor could they.

**JUSTICE BRENNAN, with whom JUSTICE MARSHALL joins, dissenting.**

If the Fourteenth Amendment is to have its intended effect as a restraint on the abuse of state power, courts must be sensitive to the manner in which state power is exercised. In an era of active government intervention to remedy social ills, the true character of the
State’s involvement in, and coercive influence over, the activities of private parties, often through complex and opaque regulatory frameworks, may not always be apparent. But if the task that the Fourteenth Amendment assigns to the courts is thus rendered more burdensome, the courts’ obligation to perform that task faithfully, and consistently with the constitutional purpose, is rendered more, not less, important.

In deciding whether “state action” is present . . . , the ultimate determination is simply whether the § 1983 defendant has brought the force of the State to bear against the § 1983 plaintiff in a manner the Fourteenth Amendment was designed to inhibit. Where the defendant is a government employee, this inquiry is relatively straightforward. But in deciding whether “state action” is present in actions performed directly by persons other than government employees, what is required is a realistic and delicate appraisal of the State’s involvement in the total context of the action taken. “Only by sifting facts and weighing circumstances can the nonobvious involvement of the State in private conduct be attributed its true significance.” The Court today departs from [this] precept, ignoring the nature of the regulatory framework presented by this case in favor of the recitation of abstract tests and a pigeonhole approach to the question of state action. But however correct the Court’s tests may be in the abstract, they are worth nothing if they are not faithfully applied. Bolstered by its own preconception of the decisionmaking process challenged by respondents, and of the relationship between the State, the nursing home operator, and the nursing home resident, the Court subjects the regulatory scheme at issue here to only the most perfunctory examination. The Court thus fails to perceive the decisive involvement of the State in the private conduct challenged by the respondents.

I

A

The Court’s analysis in this case is simple, but it is also demonstrably flawed, for it proceeds upon a premise that is factually unfounded. The Court first describes the decision to transfer a nursing home resident from one level of care to another as involving nothing more than a physician’s independent assessment of the appropriate medical treatment required by that resident. Building upon that factual premise, the Court has no difficulty concluding that the State plays no decisive role in the transfer decision . . . . If this were . . . accurate . . . , I [would agree]. A doctor who prescribes drugs for a patient on the basis of his independent medical judgment is not rendered a state actor merely because the State may reimburse the patient in different amounts depending upon which drug is prescribed.

But the level-of-care decisions at issue in this case, even when characterized as the “independent” decision of the nursing home, have far less to do with the exercise of independent professional judgment than they do with the State’s desire to save money. To be sure, standards for implementing the level-of-care scheme established by the Medicaid program are framed with reference to the underlying purpose of that program—to provide needed medical services. And not surprisingly, the State relies on doctors to implement this aspect of its Medicaid program. But the idea of two mutually exclusive levels of care—skilled nursing care and intermediate care—embodied in the federal regulatory

130

Draft — do not distribute!
scheme and implemented by the State, reflects no established medical model of health care. On the contrary, the two levels of long-term institutionalized care enshrined in the Medicaid scheme are legislative constructs, designed to serve governmental cost-containment policies.

To implement this cost-saving mechanism, the Federal Government has required States participating in the Medicaid Program to establish elaborate systems of periodic “utilization review.” With respect to patients whose expenses are not reimbursed through Medicaid, these attempts to assign the patient to one of two mutually exclusive “levels of care” would be anomalous. While the criteria used to determine which patients require the services of “skilled-nursing facilities,” which require “intermediate care facilities,” and which require no long-term institutional care at all, obviously have a medical nexus, those criteria are not geared to the specific needs of particular residents as determined by a physician; the level-of-care determination is not analogous to choosing specific medication or rehabilitative services needed by a nursing home patient. The vigor with which these reviews are performed in the nursing home context is extraordinarily unmedical in character. From a purely medical standpoint, the idea of shifting nursing home residents from a “higher level of care” to a “lower level of care,” which almost invariably involves transfer from one facility to another, rarely makes sense.

In New York, the nursing home operator is required to “maintain a discharge planning program to . . . document that the facility has made and is continuing to make all efforts possible to transfer patients to the appropriate level of care or home as indicated by the patient’s medical condition or needs” (emphasis added). The responsibility the State assigns to nursing home operators to transfer patients to appropriate levels of care is, of course, designed primarily to implement the State’s goal of reducing Medicaid costs, and the termination or reduction of benefits follows forthwith upon the facility’s discharge or transfer of a resident. “The state has, in essence, delegated a decision to . . . reduce a public assistance recipient’s benefits to a ‘private’ party” by assigning to that private party the responsibility to determine the recipient’s need. But we should not rely on that fact alone in evaluating the nexus between the State and the challenged private action. Here the State’s involvement clearly extends to supplying the standards to be used in making the delegated decision.

B

Ignoring the State’s fiscal interest in the level-of-care determination, the Court proceeds to a cursory, and misleading, discussion of the State’s involvement in the assignment of residents to particular levels of care. In my view, an accurate and realistic appraisal of the procedures actually employed in the State of New York leaves no doubt that not only has the State established the system of treatment levels and utilization review in order to further its own fiscal goals, but that the State prescribes with as much precision as is possible the standards by which individual determinations are to be made.

The Court notes that at the time of admission the admitting physician is required to complete a long-term placement form called the DMS-1. The Court dismisses the significance of the form by noting blandly that . . . “the physicians, and not the forms,
make the decision about whether the patient’s care is medically necessary. A physician can authorize a patient’s admission to a nursing facility despite a ‘low’ score on the form” (emphasis added). . . . A closer look at the regulations at issue suggests that petitioners have been less than candid in their characterization of the admission process and the role of the numerical score.

New York’s regulations mandate that the nursing home operator shall

“admit a patient only on physician’s orders and in accordance with the patient assessment criteria and standards as promulgated and published by the department (New York State LongTerm Care Placement Form [DMS-1] and New York State Numerical Standards Master Sheet [DMS-9]) . . . which shall include, as a minimum:

“(1) an assessment, performed prior to admission by or on behalf of the agency or person seeking admission for the patient, of the patient’s level of care needs according to the patient assessment criteria and standards promulgated and published by the department” (emphasis added).

The details of the DMS-9 Numerical Standards Master Sheet also bear more emphasis than the Court gives them, for that form describes with particularity the patients who are entitled to SNF care, ICF care, or no long-term residential care at all. The DMS-9 provides numerical scores for various resident [physical and mental] dysfunctions. . . .

The criterion for admission to a SNF is a DMS-9 “predictor score” of 180. For admission to an HRF (health-related facility), the required score is 60. Where the admission, or denial of admission, is based on the guidelines set forth in these regulations, there is, of course, no doubt, that the State is directly, and solely, “responsible for the specific conduct of which the plaintiff complains,” even if it has chosen to authorize a private party to implement that decision.8

The Court dismisses the specific state standards for denying admission set forth in the regulations, and tabulated according to the DMS-9, by emphasizing what it perceives as an alternative method for gaining admission to a nursing home. In the Court’s view, this alternative route to admission takes the whole scheme outside the realm of state action.

8 The Court mistakes the significance of the DMS-1, and the relevant inquiry, when it attempts to characterize that form as merely an instrument for recording the exercise of an independently exercised medical judgment. Of course, a medical background is essential in filling out the forms. But it remains clear that the State’s standards are to be applied in making the transfer determination. The Court concludes that the patient assessment standards prescribed by the State may be easily disregarded. But the regulations themselves clearly demonstrate that those standards are not merely precatory. Notably, the regulations specify that “patient assessment standards shall not be applied to residents admitted to the residential health care facility prior to March 1, 1977.” If the forms merely recorded the exercise of an independent medical judgment, rather than prescribed the standards upon which that judgment must be exercised, why would it be necessary to exempt certain patients from the inquiry? Indeed, the regulations specifically provide for a different set of standards to be applied to the continued stay review of patients admitted to a facility prior to March 1, 1977. Again, if the determination were in reality based on an independent medical assessment, it seems inconceivable to me that the State would have any interest in requiring different standards for different patients depending on when the patient had been admitted.
because it hinges on a “physician’s assessment” of what is medically necessary. In characterizing the admission process as the independent assessment of a physician, the Court relies upon, but fails to quote, the following state regulations. The language of those regulations bears noting:

“[F]or those patients failing to meet the criteria and standards for admission to the . . . facility [as measured by the DMS-9], a certification signed by a physician member of the transferring facility’s utilization review agent or signed by the responsible social services district local medicaid medical director or designee indicating the reason(s) the patient requires [the facility’s level of care, is required]” (emphasis added).

As this provision makes clear, if the potential resident does not qualify under the specific standards of the DMS-1, as tabulated on the DMS-9, the patient can be admitted only on the basis of direct approval by Medicaid officials themselves, or on the basis of a determination by the utilization review agent of the transferring facility—and, of course, such agents are themselves clearly part and parcel of the statutory cost-control process.9 No decision is made on the basis of a medical judgment exercised outside the regulatory framework, by the resident’s personal physician acting on the basis of his personal medical judgment. The attending physician’s role is, at this stage, limited to “scoring” the patient’s condition according to standards set forth by the State on the DMS-9.

Yet the State’s involvement does not end with the initial certification. Within five days after admission, the matter is again subjected to assessment, this time by the operator of the transferee facility. This time the transferee nursing home operator is required to tabulate the DMS-9 score. If the patient’s score is not adequate by the standards of the DMS-9, admission must be denied unless sanctioned by the facility’s utilization review agent.10 The utilization review agent of the admitting facility, like that of the transferring facility, operates under a “written utilization control plan, approved by the department [of health].” And that statutory body has the final say in each instance. There can thus be little doubt that in the vast majority of cases, decisions as to “level of treatment” in the admission process are made according to the State’s specified criteria. That some deviation from the most literal application of the State’s guidelines is permitted cannot change the character of the State’s involvement. Indeed, absent such provision for exceptional cases, the formularized approach embodied in the DMS-9 would be unconscionable. And indeed, even with respect to these exceptional cases, the admissions procedure is administered through bodies whose structure and operations conform to state

9 Federal regulations require each nursing home to establish a utilization review committee whose functions include review of admission decisions, and the periodic assessment of the resident’s condition to determine whether the resident’s continued stay in the facility is justified. . . .

10 A physician member of the utilization review agent has the power to determine that the patient qualifies for the type of care that the facility offers, even if the patient’s score on the DMS-1 is insufficient. If that physician member confirms that the patient is not in need of the facility’s level of care, he must then notify the patient’s attending physician “and afford that physician an opportunity for consultation.” But even if the attending physician disagrees with the adverse admission finding of the utilization review agent physician, it is the utilization review agent, not the attending physician, that makes the admission decision. The utilization review agent must, however, notify “the responsible social services district” of “any adverse admission decision.”
requirements, and whose decisions follow state guidelines—albeit guidelines somewhat more flexible than the DMS-1, in allowing some “psychosocial” factors to be taken into account.

The Court dismisses all this by noting that “[w]e cannot say that the State, by requiring completion of a form, is responsible for the physician’s decision.” The Court then notes that “[i]n any case, respondents’ complaint is about nursing home decisions to discharge or transfer, not to admit, Medicaid patients.” This is true, of course. But where, one might ask, is the Court’s discussion of the frequent utilization reviews that occur after admission? The State’s regulations require that the operator shall provide for “continued stay reviews . . . to promote efficient and effective use of available health facilities and services every 30 days for the first 90 days, and every 90 days thereafter, for each nursing home patient.”

The continued stay reviews parallel the admission determination with respect to both the State’s procedural and substantive standards.\(^{11}\) Again, the DMS-1 and the DMS-9 channel the medical inquiry and function as the principal determinants of the resident’s status, for whenever a resident does not achieve an appropriate score on the DMS-1, as determined by a nonphysician representative of the utilization review agent, the resident’s case is directed to a physician member. That physician member does not personally examine the resident, but rather relies on the DMS-1 and other documentary information. If the matter is resolved adversely to the resident, only then must the attending physician be notified. The attending physician is allowed to present relevant information, though the final decision remains with the utilization review agent. And again, the State’s substantive standards, not independent medical judgment, pervade review determinations. Evaluations are based only on the DMS-1 and DMS-9 tabulation, on a “psycho-social” evaluation respecting the resident’s response to transfer and other physical, emotional, and mental characteristics of the patient, on the resident’s discharge plan (prepared according to state regulations), and upon “additional criteria and standards . . . which shall have been approved by the department [of health]” (emphasis added).\(^{12}\)

\(^{11}\) The Court takes issue with our reliance on the nature of continued stay reviews performed by the utilization review agent, noting that “patient transfers to lower levels of care initiated by utilization review committees are simply not part of this case.” The Court’s position with respect to the work of the utilization review committee is schizophrenic at best: The Court expressly relies on its characterization of the review committee’s work as representing an independent physician’s assessment in reaching its conclusion that the DMS-1 and DMS-9 do not supply the criterion controlling the nursing home operator’s decision to admit or retain a patient in the home. In any event, the Court simply misses the point. The nursing home operator is under a continuing duty “to make all efforts possible to transfer patients to the appropriate level of care or home as indicated by the patient’s medical condition or needs.” Whether performed through the utilization review agent, or whether undertaken by the nursing home operator directly, transfers premised on the “patient’s medical condition or needs” are to be made with reference to the State’s definition of “need.”\(^{12}\)

\(^{12}\) If it is finally determined by the utilization review agent that the patient should be assigned to a lower level of care, the regulations set forth an elaborate scheme of review before the State Department of Health. These provisions apply even when the attending physician concurs in the determination. The utilization review committee must notify the Department of Health of its adverse finding and “send to the department a written statement setting forth, in specific detail, the changed medical conditions or other circumstances of the individual which support the utilization review agent’s decision for transfer, and a copy of the completed patient assessment form (DMS-1) used by the utilization review agent in this review. The 134

Draft — do not distribute!
The Court concludes with this assessment of the statutory scheme:

“These regulations do not require the nursing homes to rely on the forms in making discharge or transfer decisions, nor do they demonstrate that the State is responsible for the decision to discharge or transfer particular patients. Those decisions ultimately turn on medical judgments made by private parties according to professional standards that are not established by the State.”

The Court is wrong. As a fair reading of the relevant regulations makes clear, the State (and Federal Government) have created, and administer, the level system as a cost-saving tool of the Medicaid program. The impetus for this active program of review imposed upon the nursing home operator is primarily this fiscal concern. The State has set forth precisely the standards upon which the level-of-care determinations are to be made, and has delegated administration of the program to the nursing home operators, rather than assume the burden of administering the program itself. Thus, not only does the program implement the State’s fiscal goals, but, to paraphrase the Court, “[t]hese requirements . . . make the State responsible for actual decisions to discharge or transfer particular patients.” Where, as here, a private party acts on behalf of the State to implement state policy, his action is state action.

II

The deficiency in the Court’s analysis is dramatized by its inattention to the special characteristics of the nursing home. Quite apart from the State’s specific involvement in the transfer decisions at issue in this case, the nature of the nursing home as an institution, sustained by state and federal funds, and pervasively regulated by the State so as to ensure that it is properly implementing the governmental undertaking to provide assistance to the elderly and disabled that is embodied in the Medicaid program, undercuts the Court’s sterile approach to the state action inquiry in this case. The private nursing homes of the Nation exist, and profit, at the sufferance of state and federal Medicaid and Medicare agencies. The degree of interdependence between the State and the nursing home is far more pronounced than it was between the State and the private entity in Burton. The State subsidizes practically all of the operating and capital costs of the facility, and pays the medical expenses of more than 90% of its residents. And, in setting reimbursement rates, the State generally affords the nursing homes a profit as well. Even more striking is the fact that the residents of those homes are, by definition, utterly dependent on the State for their support and their placement. For many, the totality of their social network is the nursing home community. Within that environment, the nursing home operator is the immediate authority, the provider of food, clothing, shelter, and health care, and, in every significant respect, the functional equivalent of a State. Surely, in this context we must be especially alert to those situations in which the State department shall review the adverse continued stay finding” (emphasis added). Of course, there is no doubt that the determinations made on this review represent state action because they are performed by state officials. But if the initial determinations were not made according to state-established standards and for the State’s purposes, and were in fact “independent” medical decisions as characterized by the Court, it is difficult to understand the State’s active role in reviewing the substance of those determinations.
“has elected to place its power, property and prestige behind” the actions of the nursing home owner.

Yet, whatever might be the status of the nursing home operator where the State has simply left the resident in his charge, while paying for the resident’s support and care, it is clear that the State has not simply left nursing home patients to the care of nursing home operators. No one would doubt that nursing homes are “pervasively regulated” by State and Federal Governments; virtually every action by the operator is subject to state oversight. But the question at this stage is not whether the procedures set forth in the state and federal regulatory scheme are sufficient to protect the residents’ interests. We are confronted with the question preliminary to any Fourteenth Amendment challenge: whether the State has brought its force to bear against the plaintiffs through the office of these private parties. In answering that question we may safely assume that when the State chooses to perform its governmental undertakings through private institutions, and with the aid of private parties, not every action of those private parties is state action. But when the State directs, supports, and encourages those private parties to take specific action, that is state action.

We may hypothesize many decisions of nursing home operators that affect patients, but are not attributable to the State. But with respect to decisions to transfer patients downward from one level of care to another, if that decision is in any way connected with the statutory review structure set forth above, then there is no doubt that the standard for decision, and impetus for the decision, is the responsibility of the State. Indeed, with respect to the level-of-care determination, the State does everything but pay the nursing home operator a fixed salary. Because the State is clearly responsible for the specific conduct of petitioners about which respondents complain, and because this renders petitioners state actors for purposes of the Fourteenth Amendment, I dissent.


CHIEF JUSTICE BURGER delivered the opinion of the Court.

We [consider] whether a private school, whose income is derived primarily from public sources and which is regulated by public authorities, acted under color of state law when it discharged certain employees.

I

Respondent Kohn is the director of the New Perspectives School, a nonprofit institution located on privately owned property in Brookline, Massachusetts. The school was founded as a private institution and is operated by a board of directors, none of whom are public officials or are chosen by public officials. The school specializes in dealing with students who have experienced difficulty completing public high schools; many have drug, alcohol, or behavioral problems, or other special needs. In recent years, nearly all of the students at the school have been referred to it by the Brookline or Boston School
Committees, or by the Drug Rehabilitation Division of the Massachusetts Department of Mental Health. The school issues high school diplomas certified by the Brookline School Committee.

When students are referred to the school by Brookline or Boston under Chapter 766 of the Massachusetts Acts of 1972, the School Committees in those cities pay for the students’ education. The school also receives funds from a number of other state and federal agencies. In recent years, public funds have accounted for at least 90%, and in one year 99%, of respondent school’s operating budget. There were approximately 50 students at the school in those years and none paid tuition.

To be eligible for tuition funding under Chapter 766, the school must comply with a variety of regulations, many of which are common to all schools. The State has issued detailed regulations concerning matters ranging from recordkeeping to student-teacher ratios. Concerning personnel policies, the Chapter 766 regulations require the school to maintain written job descriptions and written statements describing personnel standards and procedures, but they impose few specific requirements.

The school is also regulated by Boston and Brookline as a result of its Chapter 766 funding. By its contract with the Boston School Committee, which refers to the school as a “contractor,” the school must agree to carry out the individualized plan developed for each student referred to the school by the Committee. The contract specifies that school employees are not city employees.

The school also has a contract with the State Drug Rehabilitation Division. Like the contract with the Boston School Committee, that agreement refers to the school as a “contractor.” It provides for reimbursement for services provided for students referred to the school by the Drug Rehabilitation Division, and includes requirements concerning the services to be provided. Except for general requirements, such as an equal employment opportunity requirement, the agreement does not cover personnel policies.

While five of the six petitioners were teachers at the school, petitioner Rendell-Baker was a vocational counselor hired under a grant from the federal Law Enforcement Assistance Administration, whose funds are distributed in Massachusetts through the State Committee on Criminal Justice. As a condition of the grant, the Committee on Criminal Justice must approve the school’s initial hiring decisions. The purpose of this requirement is to insure that the school hires vocational counselors who meet the qualifications

---

1 [Massachusetts law] requires school committees to identify students with special needs and to develop suitable educational programs for such students. [It also] provides that school committees may “enter into an agreement with any public or private school, agency, or institution to provide the necessary special education” for these students. A student identified as having special needs and recommended for placement in private school may remain in public school, if his parents object to a placement in a particular private school, unless he is especially disruptive or dangerous. Parents who object to placement in a particular private school may also elect to place their child in a private school of their choice; in such cases, they must pay the tuition.

2 . . . [Many other private schools] have a student population which is more or less evenly divided between students referred and paid for by the State and students referred and paid for by their parents or guardians.
described in the school’s grant proposal to the Committee; the Committee does not interview applicants for counselor positions.

Rendell-Baker [and the other five petitioners were] discharged by the school [because of disagreements about school policy]. . . . [T]hey sought relief under § 1983, alleging that their rights under the First, Fifth, and Fourteenth Amendments had been violated. . . .

II

A

. . . [The First and Fourteenth Amendments only apply to state action.] And § 1983, which was enacted pursuant to the authority of Congress to enforce the Fourteenth Amendment, prohibits interference with federal rights under color of state law. . . .

“In cases under § 1983, ‘under color’ of law has consistently been treated as the same thing as the ‘state action’ required under the Fourteenth Amendment.” The ultimate issue in determining whether a person is subject to suit under § 1983 is the same question posed in cases arising under the Fourteenth Amendment: is the alleged infringement of federal rights “fairly attributable to the State?” The core issue presented in this case is not whether petitioners were discharged because of their speech or without adequate procedural protections, but whether the school’s action in discharging them can fairly be seen as state action. If the action of the respondent school is not state action, our inquiry ends.

6 The Court has concluded that the acts of a private party are fairly attributable to the state on certain occasions when the private party acted in concert with state actors. For example, in Adickes v. S. H. Kress & Co. (1970), the issue was whether a restaurant violated § 1983 by refusing service to a white teacher who was in the company of six Negro students; the town sheriff arrested the white teacher for vagrancy as a result of her request to be served lunch in their company. The Court concluded that the restaurant acted under color of state law because it conspired with the sheriff, a state actor, in depriving the white teacher of federal rights.

Similarly, Flagg Brothers, Inc. v. Brooks (1978) and Lugar v. Edmondson Oil Co. (1982) illustrate the relevance of whether action was taken in concert with a state actor. The issue in Flagg Brothers was whether a warehouseman could be sued under § 1983 because it sought to execute a lien by selling goods in its possession pursuant to New York law. While the sale was authorized by a state statute, and hence appeared to be threatened under color of state law, the Court did not reach that issue. Instead, it concluded that the warehouseman’s decision to threaten to sell the goods was not “properly attributable to the State of New York,” since no state actor was involved. Since the respondent in Flagg Brothers claimed that the warehouseman violated her Fourteenth Amendment rights to due process and equal protection, and the Fourteenth Amendment is only offended by action of the state, we held that no claim for relief had been stated.

In Lugar, a lessee obtained an ex parte writ of attachment pursuant to a state statute, which was executed by a sheriff. The Court held that § 1983 applied because the involvement of the sheriff distinguished the case from Flagg Brothers. The lessee thus acted under color of state law and the sheriff’s involvement satisfied the state action requirement.

The limited role played by the Massachusetts Committee on Criminal Justice in the discharge of Rendell-Baker is not comparable to the role played by the public officials in Adickes and Lugar. The uncontradicted evidence presented by the school showed that the Committee had the power only initially to review the qualifications of a counselor selected by the school to insure that the counselor met the requirements described in the school’s grant application. The Committee had no power to hire or discharge
In *Blum*, . . . the Court considered each of the factors alleged by petitioners here to make the discharge decisions of the New Perspectives School fairly attributable to the State. . . .

The school, like the nursing homes, is not fundamentally different from many private corporations whose business depends primarily on contracts to build roads, bridges, dams, ships, or submarines for the government. Acts of such private contractors do not become acts of the government by reason of their significant or even total engagement in performing public contracts. . . .

. . . [T]he decisions to discharge the petitioners were not compelled or even influenced by any state regulation. Indeed, in contrast to the extensive regulation of the school generally, the various regulators showed relatively little interest in the school’s personnel matters. The most intrusive personnel regulation promulgated by the various government agencies was the requirement that the Committee on Criminal Justice had the power to approve persons hired as vocational counselors. Such a regulation is not sufficient to make a decision to discharge, made by private management, state action.

The third factor asserted to show that the school is a state actor is that it performs a “public function.” However, our holdings have made clear that the relevant question is not simply whether a private group is serving a “public function.” We have held that the question is whether the function performed has been “traditionally the exclusive prerogative of the State.” There can be no doubt that the education of maladjusted high school students is a public function, but that is only the beginning of the inquiry. [Massachusetts law] demonstrates that the State intends to provide services for such students at public expense. That legislative policy choice in no way makes these services the exclusive province of the State. Indeed, . . . until recently the State [has] not undertaken to provide education for students who [can] not be served by traditional public schools. That a private entity performs a function which serves the public does not make its acts state action. 7

Fourth, petitioners argue that there is a “symbiotic relationship” between the school and the State similar to the relationship involved in *Burton*. Such a claim is rejected in *Blum*, and we reject it here. . . . Here the school’s fiscal relationship with the State is not different from that of many contractors performing services for the government. No symbiotic relationship such as existed in *Burton* exists here. . . .

---

7 There is no evidence that the State has attempted to avoid its constitutional duties by a sham arrangement which attempts to disguise provision of public services as acts of private parties. Cf. *Evans v. Newton* (1966) (private trustees appointed to manage previously public park for white persons only).
JUSTICE MARSHALL, with whom JUSTICE BRENNAN joins, dissenting.

. . . The State has delegated to the New Perspectives School its statutory duty to educate children with special needs. The school receives almost all of its funds from the State, and is heavily regulated. This nexus between the school and the State is so substantial that the school’s action must be considered state action. I therefore dissent.

I

The critical facts of this case deserve restatement. [Massachusetts law] provides that all students with special needs are entitled to a suitable publicly funded education under the supervision of the state and local governments. The school committee of every city, town, or school district in Massachusetts must identify all children who, because of physical or emotional disability, have special educational needs. It must prepare an individualized educational program tailored to meet those needs, and arrange for the implementation of that program. The school committee may offer the programs through existing public schools, or it may contract with private schools to implement the programs. If the school committee decides to place a child in a private school, it must bear all the expenses associated with the placement; parents need not pay the tuition.

If a school committee decides to place a child in a private school, it must closely monitor the child’s educational progress. Every three months it must determine whether the child can be transferred to a less restrictive environment, such as a public school. In general, special education programs must be provided in the least restrictive environment possible. If the parents object to the placement of their child in private school, the child may remain in public school unless he is disruptive or dangerous. Parents may also place their child in a private school of their own choice. If they do so, however, they must pay the tuition.

As of 1978, all 50 students enrolled at the New Perspectives School were children with alcohol, drug, behavioral, or other special problems. They had been placed there pursuant to [the statute] by the town of Brookline, the city of Boston, or the Drug Rehabilitation Division of the Massachusetts Department of Mental Health. None of the students pays tuition. When they graduate, they receive a diploma certified by the Town of Brookline School Committee.

The New Perspectives School is funded almost entirely by governmental agencies. In fiscal year 1975-1976, public funds accounted for 91% of the school’s budget. In fiscal year 1976-1977, public funds accounted for 99% of the budget. The school has received money from the town of Brookline, the Massachusetts Department of Mental Health, the Massachusetts Department of Youth Services, the Massachusetts Division of Family and Children’s Services, the Massachusetts Office for Children, and the federal Law Enforcement Assistance Administration.

In order to remain eligible for placements and funding . . . , the New Perspectives School must comply with a variety of regulations. The Massachusetts Department of Education has promulgated “Guidelines for Approval of Day Educational Component in Private
Schools . . . .” These guidelines cover almost every aspect of a private school’s operations, including financial recordkeeping, student discipline, medical examinations for students, parent involvement, health care, subjects of instruction, teacher-student ratio, student records, confidentiality of records, transportation, insurance, nutrition, food preparation, toileting procedures, physical facilities, and classroom equipment. The guidelines also address personnel policies. They set forth minimum standards for staff training, use of volunteers, teacher qualifications, and teacher evaluations. They further require that the school maintain written job descriptions and a written policy on criteria and procedures for hiring and dismissal, and procedures for handling staff complaints. And they require that the school provide vacations and other benefits.

The New Perspectives School is subject to additional regulation under contracts with each of the governmental units that refers students. A contract with the Massachusetts Department of Mental Health, Drug Rehabilitation Division, requires the school to provide counseling, educational, and vocational services for drug abusers. Under a contract with the city of Boston, the school must carry out the educational plan devised by the Boston School Committee for each Boston student placed with the school. The school must submit periodic reports to the city and is subject to inspection at any time during normal business hours. Finally, the school is bound by regulations contained in contracts with the Massachusetts Department of Youth Services and the Brookline School Committee.

II

The decisions of this Court clearly establish that where there is a symbiotic relationship between the State and a privately owned enterprise, so that the State and a privately owned enterprise are participants in a joint venture, the actions of the private enterprise may be attributable to the State. “Conduct that is formally ‘private’ may become so entwined with governmental policies or so impregnated with a governmental character” that it can be regarded as governmental action. The question whether such a relationship exists “can be determined only in the framework of the peculiar facts or circumstances present.” Here, an examination of the facts and circumstances leads inexorably to the conclusion that the actions of the New Perspectives School should be attributed to the State; it is difficult to imagine a closer relationship between a government and a private enterprise.

The New Perspectives School receives virtually all of its funds from state sources. This financial dependence on the State is an important indicium of governmental involvement. The school’s very survival depends on the State. If the State chooses, it may exercise complete control over the school’s operations simply by threatening to withdraw financial support if the school takes action that it considers objectionable.

The school is heavily regulated and closely supervised by the State. . . . The school’s freedom of decisionmaking is substantially circumscribed by the Massachusetts Department of Education’s guidelines and the various contracts with state agencies. For example, the school is required to develop and comply with written rules for hiring and
dismissal of personnel. Almost every decision the school makes is substantially affected in some way by the State’s regulations.

The fact that the school is providing a substitute for public education is also an important indicium of state action. The provision of education is one of the most important tasks performed by government: it ranks at the very apex of the function of a State. Of course, as the majority emphasizes, performance of a public function is by itself sufficient to justify treating a private entity as a state actor only where the function has been “traditionally the exclusive prerogative of the State.” But the fact that a private entity is performing a vital public function, when coupled with other factors demonstrating a close connection with the State, may justify a finding of state action.

The school’s provision of a substitute for public education deserves particular emphasis because of the role of [the statute]. Under this statute, the State is required to provide a free education to all children, including those with special needs. Clearly, if the State had decided to provide the service itself, its conduct would be measured against constitutional standards. The State should not be permitted to avoid constitutional requirements simply by delegating its statutory duty to a private entity. In my view, such a delegation does not convert the performance of the duty from public to private action when the duty is specific and the private institution’s decisionmaking authority is significantly curtailed.

When an entity is not only heavily regulated and funded by the State, but also provides a service that the State is required to provide, there is a very close nexus with the State. Under these circumstances, it is entirely appropriate to treat the entity as an arm of the State. Here, since the New Perspectives School exists solely to fulfill the State’s obligations under [the statute], I think it fully reasonable to conclude that the school is a state actor.

Indeed, I would conclude that the actions challenged here were under color of state law, even if I believed that the sole basis for state action was the fact that the school was providing [services under the statute]. Petitioners claim that they were discharged because they supported student demands for increased responsibilities in school affairs, that is, because they criticized the school’s educational policies. If petitioners’ allegations are true, then the school has adopted a specific view of the sort of education that should

---

1 The majority argues that the fact that the school receives almost all of its funds from the State is not enough, by itself, to justify a finding of state action. It also contends that the fact that the school is closely supervised and heavily regulated is not enough, by itself, to justify such a finding. I am in general agreement with both propositions. However, when these two factors are present in the same case, and when other indicia of state action are also present, a finding of state action may very well be justified. By analyzing the various indicia of state action separately, without considering their cumulative impact, the majority commits a fundamental error.

2 A State may not deliberately delegate a task to a private entity in order to avoid its constitutional obligations. But a State’s decision to delegate a duty to a private entity should be carefully examined even when it has acted, not in bad faith, but for reasons of convenience. The doctrinal basis for the state action requirement is that exercises of state authority pose a special threat to constitutional values. A private entity vested with state authority poses that threat just as clearly as a state agency.
be provided under the statute, and refuses to tolerate departures from that view. The State, by refusing to intervene, has effectively endorsed that view of its duties under [the statute]. In short, because petitioners’ criticism was directly addressed to the State’s responsibilities under [the statute], a finding of state action is justified.

The majority repeatedly compares the school to a private contractor that “depends primarily on contracts to build roads, bridges, dams, ships, or submarines for the government.” The New Perspectives School can be readily distinguished, however. Although shipbuilders and dambuilders, like the school, may be dependent on government funds, they are not so closely supervised by the government. And unlike most private contractors, the school is performing a statutory duty of the State.

The majority also focuses on the fact that the actions at issue here are personnel decisions. It would apparently concede that actions directly affecting the students could be treated as under color of state law, since the school is fulfilling the State’s obligations to those children under [the statute]. It suggests, however, that the State has no interest in personnel decisions. As I have suggested, I do not share this narrow view of the school’s obligations; the personnel decisions challenged here are related to the provision of [education under the statute]. In any event, since the school is funded almost entirely by the State, is closely supervised by the State, and exists solely to perform the State’s statutory duty to educate children with special needs—since the school is really just an arm of the State—its personnel decisions may appropriately be considered state action.

III

Even though there are myriad indicia of state action in this case, the majority refuses to find that the school acted under color of state law when it discharged petitioners. The decision in this case marks a return to empty formalism in state action doctrine. Because I believe that the state action requirement must be given a more sensitive and flexible interpretation than the majority offers, I dissent.

---

4 This Court has previously emphasized the close relationship between teachers’ free speech and the educational process. The Commonwealth of Massachusetts has recently promulgated regulations recognizing that the role of teachers of special needs students is not limited to course instruction. These regulations provide:

“[T]he candidate will demonstrate that he or she:
1. responds to the needs of individual students so as to enhance their self-esteem and development
2. establishes constructive relationships with parents and others primarily concerned with the well-being of his or her students
3. works to develop a learning environment which is favorable to openness of inquiry and devoid of ridicule.”

5 In my view, this connection between the teacher’s role and the provision of Chapter 766 education would justify a finding that the State had acted under color of state law, even if the school did not depend solely on Chapter 766 placements. If the school had only one special needs student, and petitioners were discharged for criticizing the school’s education of that child, a finding of state action might be justified.

JUSTICE BLACKMUN delivered the opinion of the Court.

This case presents the question whether a physician who is under contract with the State to provide medical services to inmates at a state-prison hospital on a part-time basis acts “under color of state law,” within the meaning of 42 U.S.C. § 1983, when he treats an inmate.

I

Petitioner [West injured himself while incarcerated and was transferred to] Central Prison Hospital, the acute-care medical facility operated by the State for its more than 17,500 inmates. Central Prison Hospital has one full-time staff physician, and obtains additional medical assistance under “Contracts for Professional Services” between the State and area physicians.

Respondent, Samuel Atkins, M.D., a private physician, provided orthopedic services to inmates pursuant to one such contract. Under it, Doctor Atkins was paid approximately $52,000 annually to operate two “clinics” each week at Central Prison Hospital, with additional amounts for surgery.1 Over a period of several months, he treated West’s injury by placing his leg in a series of casts. West alleges that although the doctor acknowledged that surgery would be necessary, he refused to schedule it, and that he eventually discharged West while his ankle was still swollen and painful, and his movement still impeded. Because West was a prisoner in “close custody,” he was not free to employ or elect to see a different physician of his own choosing.2

[West sued Dr. Atkins4 under] 42 U.S.C. § 1983 . . . for violation of his Eighth Amendment right to be free from cruel and unusual punishment . . .

---

1 Doctor Atkins’ contractual duties included the following: to provide two orthopedic clinics per week; to see all orthopedic and neurological referrals; to perform orthopedic surgery as scheduled; to conduct rounds as often as necessary for his surgical and other orthopedic patients; to coordinate with the Physical Therapy Department; to request the assistance of neurosurgical consultants on spinal surgical cases; and to provide emergency on-call orthopedic services 24 hours per day. His contract required him to furnish two days of professional service each week in fulfillment of these duties. Atkins also had supervisory authority over Department of Correction nurses and physician’s assistants, who were subject to his orders.

2 North Carolina law bars all but minimum-security prisoners [(which petitioner is not)] from exercising an option to go outside the prison and obtain medical care of their choice at their own expense or funded by family resources or private health insurance.

4 Doctor Atkins is represented here by the Attorney General of North Carolina. By statute, the State provides for representation and protection from liability for any person who provides medical services to inmates and who is sued pursuant to § 1983. The State has informed its contract physicians, however, that it will not provide them with representation and indemnification in malpractice actions.

144

Draft — do not distribute!
II

To state a claim under § 1983, a plaintiff must allege the violation of a right secured by the Constitution and laws of the United States, and must show that the alleged deprivation was committed by a person acting under color of state law. Petitioner . . . sought to fulfill the first requirement by alleging a violation of his rights secured by the Eighth Amendment under Estelle v. Gamble (1976). There the Court held that deliberate indifference to a prisoner’s serious medical needs, whether by a prison doctor or a prison guard, is prohibited by the Eighth Amendment. . . .

A

The traditional definition of acting under color of state law requires that the defendant in a § 1983 action have exercised power “possessed by virtue of state law and made possible only because the wrongdoer is clothed with the authority of state law.” In Lugar, the Court made clear that if a defendant’s conduct satisfies the state-action requirement of the Fourteenth Amendment, “that conduct [is] also action under color of state law and will support a suit under § 1983.” In such circumstances, the defendant’s alleged infringement of the plaintiff’s federal rights is “fairly attributable to the State.”

To constitute state action, “the deprivation must be caused by the exercise of some right or privilege created by the State . . . or by a person for whom the State is responsible,” and “the party charged with the deprivation must be a person who may fairly be said to be a state actor.” “[S]tate employment is generally sufficient to render the defendant a state actor.” It is firmly established that a defendant in a § 1983 suit acts under color of state law when he abuses the position given to him by the State. Thus, generally, a public employee acts under color of state law while acting in his official capacity or while exercising his responsibilities pursuant to state law.

Indeed, Polk County v. Dodson (1981) . . . is the only case in which this Court has determined that a person who is employed by the State and who is sued under § 1983 for abusing his position in the performance of his assigned tasks was not acting under color of state law. The Court held that “a public defender does not act under color of state law when performing a lawyer’s traditional functions as counsel to a defendant in a criminal proceeding.” In this capacity, the Court noted, a public defender differs from the typical government employee and state actor. While performing his duties, the public defender retains all of the essential attributes of a private attorney, including, most importantly, his “professional independence,” which the State is constitutionally obliged to respect. A criminal lawyer’s professional and ethical obligations require him to act in a role independent of and in opposition to the State. The Court accordingly concluded that when representing an indigent defendant in a state criminal proceeding, the public defender does not act under color of state law for purposes of § 1983 because he “is not acting on behalf of the State; he is the State’s adversary.”
. . . In contrast to the public defender, Doctor Atkins’ professional and ethical obligation to make independent medical judgments did not set him in conflict with the State and other prison authorities. Indeed, his relationship with other prison authorities was cooperative. . . .

. . . [P]rofessionals [may] act under color of state law [even] when they act in their professional capacities. . . . [A] professional [need not] . . . exercis[e] “custodial or supervisory” authority [to be subject to suit under § 1983]. To the extent this Court in *Polk County* relied on the fact that the public defender is a “professional” in concluding that he was not engaged in state action, the case turned on the particular professional obligation of the criminal defense attorney to be an adversary of the State, not on the independence and integrity generally applicable to professionals as a class. Indeed, . . . [in previous cases,] this Court either has identified professionals as state actors, see, *e.g.*, *Tower v. Glover* (1984) (state public defenders), or has assumed that professionals are state actors in § 1983 suits, see, *e.g.*, *Estelle v. Gamble* (1976) (medical director of state prison who was also the treating physician). Defendants are not removed from the purview of § 1983 simply because they are professionals acting in accordance with professional discretion and judgment.¹⁰

[Moreover,] . . . this Court’s decision in *Estelle* . . . demonstrates that custodial and supervisory functions are irrelevant to an assessment whether the particular action challenged was performed under color of state law. In *Estelle*, the inmate’s Eighth Amendment claim was brought against the physician-employee, Dr. Gray, in his capacity both as treating physician and as medical director of the state prison system. Gray was sued, however, solely on the basis of allegedly substandard medical treatment given to the plaintiff; his supervisory and custodial functions were not at issue. . . . [T]he inference to be drawn from *Estelle* is that the medical treatment of prison inmates by prison physicians is state action. The Court explicitly held that “indifference . . . manifested by

¹⁰ We do not suggest that this factor is entirely irrelevant to the state-action inquiry. Where the issue is whether a *private* party is engaged in activity that constitutes state action, it may be relevant that the challenged activity turned on judgments controlled by professional standards, where those standards are not established by the State. The Court has held that “a State normally can be held responsible for a private decision only when it has exercised coercive power or has provided such significant encouragement, either overt or covert, that the choice must in law be deemed to be that of the State.” In both *Blum v. Yaretsky* and *Rendell-Baker*, the fact that the private entities received state funding and were subject to state regulation did not, without more, convert their conduct into state action. The Court suggested that the private party’s challenged decisions could satisfy the state-action requirement if they were made on the basis of some rule of decision for which the State is responsible. The Court found, however, that the decisions were based on independent professional judgments and were not subject to state direction. Thus, the requisite “nexus” to the State was absent.

This determination cannot be transformed into the proposition that no person acts under color of state law where he is exercising independent professional judgment. . . . *Blum* and *Rendell-Baker* provide no support for respondent’s argument that a physician, employed by the State to fulfill the State’s constitutional obligations, does not act under color of state law merely because he renders medical care in accordance with professional obligations.
prison doctors in their response to the prisoner’s needs . . . states a cause of action under § 1983.” . . .

C

We now make explicit what was implicit in our holding in *Estelle*: Respondent, as a physician employed by North Carolina to provide medical services to state prison inmates, acted under color of state law for purposes of § 1983 when undertaking his duties in treating petitioner’s injury. Such conduct is fairly attributable to the State.

The Court recognized in *Estelle*: “An inmate must rely on prison authorities to treat his medical needs; if the authorities fail to do so, those needs will not be met.” In light of this, the Court held that the State has a constitutional obligation, under the Eighth Amendment, to provide adequate medical care to those whom it has incarcerated. North Carolina employs physicians, such as respondent, and defers to their professional judgment, in order to fulfill this obligation. By virtue of this relationship, effected by state law, Doctor Atkins is authorized and obliged to treat prison inmates, such as West. He does so “clothed with the authority of state law.” He is “a person who may fairly be said to be a state actor.” It is only those physicians authorized by the State to whom the inmate may turn. Under state law, the only medical care West could receive for his injury was that provided by the State. If Doctor Atkins misused his power by demonstrating deliberate indifference to West’s serious medical needs, the resultant deprivation was caused, in the sense relevant for state-action inquiry, by the State’s exercise of its right to punish West by incarceration and to deny him a venue independent of the State to obtain needed medical care.

The fact that the State employed respondent pursuant to a contractual arrangement that did not generate the same benefits or obligations applicable to other “state employees” does not alter the analysis. It is the physician’s function within the state system, not the precise terms of his employment, that determines whether his actions can fairly be attributed to the State. Whether a physician is on the state payroll or is paid by contract, the dispositive issue concerns the relationship among the State, the physician, and the prisoner. Contracting out prison medical care does not relieve the State of its constitutional duty to provide adequate medical treatment to those in its custody, and it does not deprive the State’s prisoners of the means to vindicate their Eighth Amendment rights. The State bore an affirmative obligation to provide adequate medical care to West; the State delegated that function to respondent Atkins; and respondent voluntarily assumed that obligation by contract.

Nor does the fact that Doctor Atkins’ employment contract did not require him to work exclusively for the prison make him any less a state actor than if he performed those duties as a full-time, permanent member of the state prison medical staff. It is the physician’s function while working for the State, not the amount of time he spends in performance of those duties or the fact that he may be employed by others to perform
similar duties, that determines whether he is acting under color of state law. In the State’s employ, respondent worked as a physician at the prison hospital fully vested with state authority to fulfill essential aspects of the duty, placed on the State by the Eighth Amendment and state law, to provide essential medical care to those the State had incarcerated. Doctor Atkins must be considered to be a state actor.

§ 2.10. Notes on the “quasi-governmental organization” cases

1. Different lines of state action cases. These materials distinguish between two lines of cases. First, there are the “individual remedies” cases like Flagg Bros. and Lugar, where individual creditors or litigants use state-authorized procedures to seize (what seems like) other people’s property. Second, there are the “quasi-governmental organization” cases like Rendell-Baker, Blum, Sullivan, and Brentwood Academy, where the question is whether an organization that may seem private, like a school or nursing home, is in fact “related” enough to the government that (at least some of) its action should be described as state action. The criterion of “relatedness” can be described as a “nexus,” “entwinement,” regulatedness, etc., depending on the case. Keep in mind that this distinction is purely a matter of organizational convenience and doesn’t correspond to any bright lines or possibly even principled distinctions.

2. Government compulsion or influence. Both Blum and Rendell-Baker recite the doctrinal test of whether the private decision was compelled or influenced by state regulation. For an example of where such influence was found, see the lunch counter sit-in cases.

In Peterson v. City of Greenville (1963), the manager of an S.H. Kress store refused to serve blacks at his lunch counter and called the police, who arrested them for trespass. The city of Greenville had an ordinance requiring racial separation in restaurants. Chief Justice Warren wrote: “When the State has commanded a particular result, it has saved to itself the power to determine that result and thereby ‘to a significant extent’ has ‘become involved’ in it, and, in fact, has removed that decision from the sphere of private choice.” This was true even if the manager would have discriminated anyway. The trespass convictions were overturned, but not on the Shelley rationale that it was unconstitutional for a court to enforce neutral laws based on private discrimination, but on the theory that the manager himself was a state actor and the unconstitutional discrimination occurred at the moment that the manager obeyed the law and refused service.

15 Contrary to respondent’s intimations, the fact that a state employee’s role parallels one in the private sector is not, by itself, reason to conclude that the former is not acting under color of state law in performing his duties. “If an individual is possessed of state authority and purports to act under that authority, his action is state action. It is irrelevant that he might have taken the same action had he acted in a purely private capacity . . . .”
In a companion case, *Lombard v. Louisiana* (1963), there was no statute or ordinance compelling the discrimination. However, New Orleans city officials had issued statements earlier condemning sit-ins, announcing that the police would not permit sit-ins, and committing themselves to enforcement of the relevant breach-of-the-peace laws. Though a restaurateur could serve blacks without any penalties, Chief Justice Warren wrote that the state’s “official command” had “at least as much coercive effect as an ordinance.” The parallel to *Peterson* (and failure to cite *Shelley*) suggests (though it does not so hold directly) that again, it was the store manager’s decision, quasi-coerced by the government, that violated Equal Protection. See also Brennan’s dissent in *Adickes*. Does this theory make sense?

In *Robinson v. Florida* (1964), a restaurant refused to serve blacks in the presence of Florida “health” regulations that required integrated restaurants to provide separate toilet facilities. Based on *Peterson*, Justice Black wrote: “While these Florida regulations do not directly and expressly forbid restaurants to serve both white and colored people together, they certainly embody a state policy putting burdens upon any restaurant which serves both races, burdens bound to discourage the serving of the two races together. . . . [T]he State through its regulations has become involved to such a significant extent in bringing about restaurant segregation that appellants’ trespass convictions must be held to reflect that state policy and therefore to violate the Fourteenth Amendment.” How broad is the *Robinson* reasoning? Does a private action become state action whenever state regulation burdens the opposite choice to any extent? Any private actions in an area affected by regulations (is that everywhere?) either go in the direction burdened by regulation or in the opposite direction; and we’d expect that, if the burdens have incentive effects, one would be less likely to take an action when it is burdened by regulation. Could this make more than half of all actions state action?

Brennan, in his *Adickes* dissent, suggests that if a state statute merely authorizes discrimination (“Every person . . . is hereby authorized and empowered to refuse to sell to . . . any person that the owner . . . does not desire to sell to . . . .”), this should be enough to establish state action by the discriminator. The doctrine has not gone that far.

You should read Brennan’s dissent in *Blum*—which argues that, given state regulations on nursing homes, the state “directs, supports, and encourages” the transfer decisions—in light of these principles. There can be a fine line between authorizing and encouraging, and between encouraging and compelling. This is why it is important to look at the fine details of the regulatory schemes. Whose description of the nursing home scheme do you agree with? Brennan concedes that “[a] doctor who prescribes drugs for a patient on the basis of his independent medical judgment is not rendered a state actor merely because the State may reimburse the patient in different amounts depending upon which drug is prescribed.” Would it make a difference if the state were reimbursing the doctor
instead of the patient? Would it make a difference if the doctor were motivated by the reimbursement? Would this be consistent with a robust understanding of encouragement in line with Robinson? Is “encouragement” a sensible or coherent test?

3. **Alternate characterizations of service-provision privatization.** One can start out describing the service provided as public. For instance, Marshall’s dissent in *Rendell-Baker* describes the education provided as the state’s statutory responsibility. If the state were provided the education directly, it would clearly be state action. (But does the statute require direct provision?) Then Marshall describes the privatization as delegation to the private school of this responsibility. Seen from this direction, it seems attractive to argue that “[t]he State should not be permitted to avoid constitutional requirements simply by delegating its statutory duty to a private entity.” (Brennan’s dissent in *Blum* does not take this approach, because Brennan rests on the strong encouragement of cost-based transfer decisions by state regulations. But he could have taken this approach: first, consider state-run nursing homes; then, imagine that the provision of nursing home services is delegated to private parties.) This characterization explains Gillian Metzger’s statement that “the only reason to consider applying the Constitution outside of formal government institutions is to prevent the public-private divide from eviscerating the fundamental requirement of constitutional accountability.” Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1422 (2003).

But suppose we look at the services from the other direction. The state is under no *constitutional* duty to provide education or nursing homes (and, in *Blum*, the majority notes that the state would not necessarily provide the service by itself). Therefore, the constitutional baseline can be seen as total government withdrawal from the activity, with only private schools or nursing homes (whether for-profit or non-profit) making hiring and transfer decisions, subject only to the terms of their contracts with their teachers or patients. Now imagine that the government decides to support these worthwhile efforts by subsidizing their providers with tax money. This arguably doesn’t make the subsidy recipients into state actors—even if their money comes entirely from the state. (Suppose you want to dispute this. Discuss the status of a restaurant on Capitol Hill that does all its business catering congressional receptions.)

Suppose now that the government decides to attach restrictions on its funding, perhaps to save money and perhaps just for general regulatory purposes. (Or suppose these restrictions aren’t attached to the funding, but are direct regulations of the providers, enforced by fines or other penalties.) The funding recipients may be able to challenge the restrictions based on the unconstitutional conditions doctrine, but this is a much looser doctrine; and in any event, the existence of the restrictions or regulations doesn’t seem to make the recipients into state actors, unless (on a *Peterson* theory) the regulations explicitly mandate the challenged actions. Of course the funding recipients may change their behavior because of the restrictions (we expect that any rules have incentive effects); should this alter
the analysis? Consider note 2 supra on the Robinson rationale, how broadly to read “encouragement,” and to what extent such a test makes sense.

Try characterizing various “quasi-governmental services” cases in these alternate ways. (See also Sullivan and Brentwood Academy below.) Does the existence of a statutory responsibility (as Marshall stressed in his Rendell-Baker dissent) affect the plausibility of the recharacterization? (If so, is it right that statutory responsibilities affect constitutional duties? Should it make a difference whether the statutory responsibility is described as a specific responsibility to provide the service or an individual entitlement to obtain the service?) Does the recharacterization work for West? How about Logiodice (described in note 4 infra)?

4. **What is government power?** The potential recharacterization of government service provision described above highlights the elusiveness of the concept of “government power.” Thus, Metzger writes: “The complaint that current doctrine represents a very poor gauge of government power necessarily goes beyond an internal critique of current Supreme Court doctrine, in that it presupposes an understanding of what constitutes the exercise of such power that differs from the view evidenced in the Court’s state action decisions. Yet, it does not require a very robust or expansive understanding of government power in order to make the point that current state action doctrine is underinclusive. Instead, all that is needed is acceptance of the claim that control over third parties’ access to government resources, specifically government benefits and government-subsidized services, represents government power. The Court’s state action decisions demonstrate that current doctrine is largely unconcerned with government power in this form or, more generally, with the control over third parties that private entities gain as a result of their roles in government programs.” Metzger, at 1422.

Metzger points to Blum as an example of this underinclusiveness, where the parties making transfer decisions were not considered state actors though they controlled patients’ access to government resources. Is this characterization open-and-shut? How does this differ from a doctor prescribing one drug rather than another because he doesn’t believe he would be reimbursed for the more expensive one? Suppose we accept the view expressed in Brennan’s dissent, that the regulations were so micromanaging that they all but determined the transfer decision. Are the private parties (acting with significant help from the regulations) determining the patients’ access to government resources, or are the nursing homes merely determining the patients’ access to their own resources based on their concern with their own access to government reimbursement?

Metzger also points to San Francisco Arts & Athletics, Inc. (SFAA) v. United States Olympic Committee (USOC) (1987), where “the Court rejected the claim that the USOC was a state actor, holding that oversight of amateur sports was not a public function and that the government’s involvement in chartering, authorizing, and funding the USOC did not suffice to create state action. Notably, in reaching this conclusion the Court ignored the way that Congress’ grant to the
USOC of exclusive oversight powers significantly enhanced the USOC’s authority over all individuals and organizations participating in amateur sports and enabled it to perform a coordination role that prior consensual arrangements had failed to achieve.” Does the conclusion that the USOC exercises “authority” depend on the international rules regarding who is allowed to participate in the Olympics?

More generally, Metzger faults the Court for focusing on the extent of government involvement with a private action (a private act can be considered state action when the nexus, encouragement, entwinement, etc., is great enough) and for ignoring the nature of the act being challenged (whether the act involves “authority” over others or control over access to government resources). In fact, private parties’ authority over others can be greater when the extent of government oversight is less, which leads to a perverse underinclusiveness.

See, for instance, Logiodice v. Trustees of Maine Central Institute (1st Cir. 2002), where, in a suit by a student challenging his suspension on Due Process grounds, Judge Boudin held that a private school serving as the high school for a Maine school district was not a state actor. Boudin relied heavily on the state’s aloofness from the day-to-day operations of the school. But doesn’t the state’s lack of control enhance the school’s “potential for abusive action”? Metzger suggests that state action doctrine has been “refocused . . . away from ensuring constitutional accountability of government power and towards ensuring constitutional accountability of government proper. That is, the Court uses state action doctrine to police against intentional evasion and bad faith by those who are indisputably government actors, but it does not view the doctrine as a safeguard against private actors wielding government power outside of constitutional constraints.” But, she argues, current doctrine is not even good at policing government itself, since government can influence its private partners in many ways short of direct involvement (this is basically the concern about “encouragement” discussed above). Metzger, at 1425.

Logiodice himself was an 11th-grader and therefore perhaps below 17, which is the age up to which education is compulsory in Maine. Should this make a difference in whether you agree with Metzger that the private school’s power was government power? Does it make a difference in whether the Maine delegation can be recharacterized as private action as described in note 3 supra?

5. Overinclusiveness. Metzger writes: “At the same time, however, current state action doctrine is also overinclusive. Under standard analysis, private entities are held fully subject to constitutional constraints if they are found to be state actors, even when imposing such liability on private actors is not needed to preserve constitutional accountability.”

For example, consider Grijalva v. Shalala (9th Cir. 1998), vacated in light of the Supreme Court’s decision in Am. Mfrs. Mutual Ins. Co. v. Sullivan (1999). Under the Medicare statute, beneficiaries may enroll in a managed care organization
(MCO), which provide all covered Medicare services in exchange for a flat fee from the government. Certain beneficiaries sued, charging that they were denied certain services without adequate notice and appeal rights. The Ninth Circuit found that the MCO was a state actor, and was therefore subject to Due Process requirements. It upheld the district court’s injunction: “[T]he notice must be legible (at least 12-point type), state clearly the reason for the denial, inform the enrollee of all appeal rights, explain hearing rights and procedures, and provide ‘instruction on how to obtain supporting evidence, including medical records and supporting affidavits from the attending physician.’ . . . [A]ny hearing must be ‘informal, in-person communication with the decisionmaker,’ available upon request for all service denials, and timely.” 152 F.3d 1115, 1119, quoting 946 F. Supp. 747, 760–61 (D. Ariz. 1996).

Metzger finds this excessive: “From a constitutional accountability perspective, holding the MCOs to be state actors seems unnecessary and purely the result of doctrinal rules. The plaintiffs sued only the Secretary of Health and Human Services (HHS), not the MCOs, their goal being to force the Secretary to do a better job at policing the procedures used by MCOs in making coverage decisions. The Secretary, in turn, had disclaimed responsibility for the MCOs’ actions, arguing that they were unrelated private entities. State action doctrine thus served as a bridge that allowed the federal government to be charged with responsibility for the MCOs’ actions even though the government did not directly participate in specific coverage denials until the appeals level. Surely, however, making the leap from private to public responsibility in this type of situation should not be difficult, given that the government had given the MCOs responsibility for providing services under Medicare notwithstanding the MCOs’ inadequate procedures.

“More importantly, once that leap is made and the government is liable for private acts, constitutional accountability is assured. Extending constitutional constraints to the private entities or individuals involved gives plaintiffs additional defendants, but not additional substantive protection against constitutional violations. Such an extension, however, severely intrudes upon the government’s ability to tailor relationships with private partners so as to maximize the benefits of privatization. Instead, as Grijalva demonstrates, regulatory control is transferred to the courts. . . . [T]he net effect [of the district court’s detailed order] was to prevent the government from determining how best to achieve the constitutional demands of due process without undermining the feasibility and benefits of MCO participation in Medicare. Given the federal courts’ lack of expertise in the complex policy issues presented by managed care, responsibility for designing constitutionally adequate MCO procedures seems best left in the first instance to HHS, subject to judicial review.” Metzger, at 1426–28.

Metzger argues that this overinclusivity is not just unnecessary, since the government could be sued directly, but also harmful to constitutional rights, as courts are reluctant to find state action when such a finding would have such a
severe all-or-nothing effect. See also *Brentwood Academy*, where the Court found state action without examining whether the alternative accountability mechanisms identified by the Association were sufficient, and *Blum*, where the Court failed to find state action based on the presence of independent professional standards, without examining whether these standards were sufficient.

6. **Perverse incentives.** The section on economics and public law later in these readings discusses various considerations relevant to the choice whether to privatize, and to what extent. The withdrawal of government could be complete, or the government could retain a broad supervisory role; there are many possible points on the continuum, any one of which may be optimal given the relative importance of the need for flexibility and the danger of private abuse. However, as *Blum*, *Rendell-Baker*, and later cases show, state action is an on-off switch, which turns on when the involvement of government in the day-to-day operations of the private entity is great enough. Moreover, a finding of state action—with its requirements of due process and the like—can be highly burdensome to the private party. Does this create an incentive for the government to withdraw itself more than is optimal, in cases where one would otherwise favor some intermediate level of government involvement?

Metzger writes: “By creating such an incentive to privatize without close supervision—and a corresponding disincentive for governments to adopt cooperative models of privatization, where programs are jointly implemented by private and public entities—current state action doctrine raises seemingly indefensible obstacles to effective and accountable use of privatization. Several scholars studying privatization emphasize the importance of close government involvement and cooperation with private entities. Such involvement allows the government to benefit from private expertise and innovation, while preserving public control of government programs and guarding against self-interested private decisionmaking. Close oversight is particularly important when market failure or abuse of power is most likely: where providers hold a monopoly on provision of particular services; the services at issue are complex and difficult to specify ab initio; competitive pressures are minimized by difficulties in exit or lack of information; and recipients are relatively powerless. Even more relevant to the discussion here, close government involvement is the best means for ensuring constitutional accountability. Agency officials are likely to be more familiar with constitutional requirements than are private entities, and they also may be more concerned with acting in accordance with those requirements. And perhaps most importantly, the presence of agency officials working in partnership with private providers or regulators minimizes the difficulty in enforcing constitutional obligations, since then no need exists to invoke the state action doctrine against private entities; instead, constitutional claims can be brought against the government officials involved.” Metzger, at 1436–37.

Non-state actors can also avoid being subject to administrative law constraints, which only apply to “agency action.” Can this perverse incentive work in the other direction as well? What if greater withdrawal of government is optimal, but
the government is unwilling to cede that much control. Does state action doctrine give the government an incentive to increase its level of micromanagement to such a point that the private party becomes subject to constitutional requirements? Moreover, cases like *Blum* show that, even if the government does not manage day-to-day activities, it can still exercise effective control through more subtle encouragement, incentives, and soft accountability mechanisms. Does that suggest that the dangers of excessive cession of day-to-day management are overblown?

7. **How about non-constitutional reforms?** Of course, the state action doctrine is not the only way of guaranteeing the accountability of private organizations involved in government service provision. Non-constitutional measures include statutory or contractual appeal and hearing rights, community participation, private-law remedies, restrictions on funding, and so on. These are discussed more in the section on economics and public law later in these readings.

Metzger argues that constitutional remedies are still important because they do not depend on the good will of government—and one cannot count on the government's willingness to subject private parties to burdensome requirements when the very reason it privatized was to save money. Or, she argues, when the government does protect rights within a program, it will do so selectively and without providing mechanisms for *individual* enforcement. Moreover, to the extent state action doctrine provides perverse incentives for the government to cede too much power to private parties, constitutional doctrine reform is necessary just to facilitate the non-constitutional approach. Finally, Metzger argues that constitutional accountability is necessary on non-instrumentalist grounds, just to protect “the principle of constitutional supremacy and the legitimacy of our constitutional system.” Metzger, at 1452–56.

---

§ 2.11. **DeShaney v. Winnebago County Dep’t of Soc. Servs., 489 U.S. 189 (1989)**

**CHIEF JUSTICE REHNQUIST** delivered the opinion of the Court.

Petitioner is a boy who was beaten and permanently injured by his father, with whom he lived. Respondents are social workers and other local officials who received complaints that petitioner was being abused by his father and had reason to believe that this was the case, but nonetheless did not act to remove petitioner from his father’s custody. Petitioner sued respondents claiming that their failure to act deprived him of his liberty in violation of the Due Process Clause . . . . We hold that it did not.

I

. . . [Local] authorities first learned that Joshua DeShaney might be a victim of child abuse in January 1982, when his father’s second wife complained to the police, at the
time of their divorce, that he had previously “hit the boy causing marks and [was] a prime case for child abuse.” The . . . Department of Social Services (DSS) interviewed the father, but he denied the accusations, and DSS did not pursue them further. In January 1983, Joshua was admitted to a local hospital with multiple bruises and abrasions. The examining physician suspected child abuse and notified DSS, which immediately obtained an order from a . . . juvenile court placing Joshua in the temporary custody of the hospital. Three days later, the county convened an ad hoc “Child Protection Team”—consisting of a pediatrician, a psychologist, a police detective, the county’s lawyer, several DSS caseworkers, and various hospital personnel—to consider Joshua’s situation. At this meeting, the Team decided that there was insufficient evidence of child abuse to retain Joshua in the custody of the court. The Team did, however, decide to recommend several measures to protect Joshua, including enrolling him in a preschool program, providing his father with certain counselling services, and encouraging his father’s girlfriend to move out of the home. [His father] entered into a voluntary agreement with DSS in which he promised to cooperate with them in accomplishing these goals.

Based on the recommendation of the Child Protection Team, the juvenile court dismissed the child protection case and returned Joshua to the custody of his father. A month later, emergency room personnel called the DSS caseworker handling Joshua’s case to report that he had once again been treated for suspicious injuries. The caseworker concluded that there was no basis for action. For the next six months, the caseworker made monthly visits to the DeShaney home, during which she observed a number of suspicious injuries on Joshua’s head; she also noticed that he had not been enrolled in school, and that the girlfriend had not moved out. The caseworker dutifully recorded these incidents in her files, along with her continuing suspicions that someone in the DeShaney household was physically abusing Joshua, but she did nothing more. In November 1983, the emergency room notified DSS that Joshua had been treated once again for injuries that they believed to be caused by child abuse. On the caseworker’s next two visits to the DeShaney home, she was told that Joshua was too ill to see her. Still DSS took no action.

In March 1984, [his father] beat 4-year-old Joshua so severely that he fell into a life-threatening coma. Emergency brain surgery revealed a series of hemorrhages caused by traumatic injuries to the head inflicted over a long period of time. Joshua did not die, but he suffered brain damage so severe that he is expected to spend the rest of his life confined to an institution for the profoundly retarded. [His father] was subsequently tried and convicted of child abuse.

Joshua and his mother brought this action under 42 U.S.C. § 1983 . . . against [the county government, and DSS and some of its employees,] alleg[ing] that respondents had deprived Joshua of his liberty without due process of law, in violation of his rights under the Fourteenth Amendment, by failing to intervene to protect him against a risk of violence at his father’s hands of which they knew or should have known. . . .

II

The Due Process Clause . . . provides that “[n]o State shall . . . deprive any person of life, liberty, or property, without due process of law.” Petitioners contend that the State
deprived Joshua of his liberty interest in “free[dom] from . . . unjustified intrusions on personal security” by failing to provide him with adequate protection against his father’s violence.

But nothing in the language of the Due Process Clause itself requires the State to protect the life, liberty, and property of its citizens against invasion by private actors. The Clause is phrased as a limitation on the State’s power to act, not as a guarantee of certain minimal levels of safety and security. It forbids the State itself to deprive individuals of life, liberty, or property without “due process of law,” but its language cannot fairly be extended to impose an affirmative obligation on the State to ensure that those interests do not come to harm through other means. Nor does history support such an expansive reading of the constitutional text. . . . [T]he Due Process Clause . . . was intended to prevent government “from abusing [its] power, or employing it as an instrument of oppression.” Its purpose was to protect the people from the State, not to ensure that the State protected them from each other. The Framers were content to leave the extent of governmental obligation in the latter area to the democratic political processes.

Consistent with these principles, our cases have recognized that the Due Process Clauses generally confer no affirmative right to governmental aid, even where such aid may be necessary to secure life, liberty, or property interests of which the government itself may not deprive the individual. . . . If the Due Process Clause does not require the State to provide its citizens with particular protective services, it follows that the State cannot be held liable under the Clause for injuries that could have been averted had it chosen to provide them. As a general matter, then, we conclude that a State’s failure to protect an individual against private violence simply does not constitute a violation of the Due Process Clause.

Petitioners contend, however, that even if the Due Process Clause imposes no affirmative obligation on the State to provide the general public with adequate protective services, such a duty may arise out of certain “special relationships” created or assumed by the State with respect to particular individuals. Petitioners argue that such a “special relationship” existed here because the State knew that Joshua faced a special danger of abuse at his father’s hands, and specifically proclaimed, by word and by deed, its intention to protect him against that danger. Having actually undertaken to protect Joshua from this danger—which petitioners concede the State played no part in creating—the State acquired an affirmative “duty,” enforceable through the Due Process Clause, to do so in a reasonably competent fashion. Its failure to discharge that duty, so the argument goes, was an abuse of governmental power that so “shocks the conscience” as to constitute a substantive due process violation.

We reject this argument. It is true that in certain limited circumstances the Constitution imposes upon the State affirmative duties of care and protection with respect to particular individuals. . . .

[For instance,] when the State takes a person into its custody and holds him there against his will, the Constitution imposes upon it a corresponding duty to assume some responsibility for his safety and general well-being. The rationale for this principle is
simple enough: when the State by the affirmative exercise of its power so restrains an individual’s liberty that it renders him unable to care for himself, and at the same time fails to provide for his basic human needs—e.g., food, clothing, shelter, medical care, and reasonable safety—it transgresses the substantive limits on state action set by the Eighth Amendment and the Due Process Clause. The affirmative duty to protect arises not from the State’s knowledge of the individual’s predicament or from its expressions of intent to help him, but from the limitation which it has imposed on his freedom to act on his own behalf. In the substantive due process analysis, it is the State’s affirmative act of restraining the individual’s freedom to act on his own behalf—through incarceration, institutionalization, or other similar restraint of personal liberty—which is the “deprivation of liberty” triggering the protections of the Due Process Clause, not its failure to act to protect his liberty interests against harms inflicted by other means.

[This] analysis simply has no applicability in the present case. Petitioners concede that the harms Joshua suffered occurred not while he was in the State’s custody, but while he was in the custody of his natural father, who was in no sense a state actor. While the State may have been aware of the dangers that Joshua faced in the free world, it played no part in their creation, nor did it do anything to render him any more vulnerable to them. . . . Under these circumstances, the State had no constitutional duty to protect Joshua.

It may well be that, by voluntarily undertaking to protect Joshua against a danger it concededly played no part in creating, the State acquired a duty under state tort law to provide him with adequate protection against that danger. But the claim here is based on the Due Process Clause . . . , which . . . does not transform every tort committed by a state actor into a constitutional violation. . . . Because . . . the State had no constitutional duty to protect Joshua against his father’s violence, its failure to do so—though calamitous in hindsight—simply does not constitute a violation of the Due Process Clause.

Judges and lawyers, like other humans, are moved by natural sympathy in a case like this to find a way for Joshua and his mother to receive adequate compensation for the grievous harm inflicted upon them. But before yielding to that impulse, it is well to remember once again that the harm was inflicted not by the State of Wisconsin, but by Joshua’s father. The most that can be said of the state functionaries in this case is that they stood by and did nothing when suspicious circumstances dictated a more active role for them. In defense of them it must also be said that had they moved too soon to take custody of the son away from the father, they would likely have been met with charges of improperly intruding into the parent-child relationship, charges based on the same Due Process Clause that forms the basis for the present charge of failure to provide adequate protection.

9 Had the State by the affirmative exercise of its power removed Joshua from free society and placed him in a foster home operated by its agents, we might have a situation sufficiently analogous to incarceration or institutionalization to give rise to an affirmative duty to protect. Indeed, several Courts of Appeals have held . . . that the State may be held liable under the Due Process Clause for failing to protect children in foster homes from mistreatment at the hands of their foster parents. We express no view on the validity of this analogy, however, as it is not before us in the present case.

Draft — do not distribute!
The people of Wisconsin may well prefer a system of liability which would place upon the State and its officials the responsibility for failure to act in situations such as the present one. They may create such a system, if they do not have it already, by changing the tort law of the State in accordance with the regular lawmaking process. But they should not have it thrust upon them by this Court’s expansion of the Due Process Clause of the Fourteenth Amendment.

**JUSTICE BRENNAN, with whom JUSTICE MARSHALL and JUSTICE BLACKMUN join, dissenting.**

“The most that can be said of the state functionaries in this case,” the Court today concludes, “is that they stood by and did nothing when suspicious circumstances dictated a more active role for them.” Because I believe that this description of respondents’ conduct tells only part of the story and that, accordingly, the Constitution itself “dictated a more active role” for respondents in the circumstances presented here, I cannot agree that respondents had no constitutional duty to help Joshua DeShaney.

[Rather than] leading off with a discussion (and rejection) of the idea that the Constitution imposes on the States an affirmative duty to take basic care of their citizens[—which no one ever raised in this case—] . . . I would focus first on the action that Wisconsin has taken with respect to Joshua and children like him . . . . [The incarceration and institutionalization cases] began by emphasizing that the States had confined [the affected litigants, leaving them] helpless to help themselves or to seek help from persons unconnected to the government. . . . [M]oreover, actual physical restraint is not the only state action that has been considered relevant. See, e.g., *White v. Rochford* (7th Cir. 1979) (police officers violated due process when, after arresting the guardian of three young children, they abandoned the children on a busy stretch of highway at night).

. . . [T]o the Court, the only fact that seems to count as an “affirmative act of restraining the individual’s freedom to act on his own behalf” is direct physical control. I would not, however, give [the incarceration and institutionalization cases] such a stingy scope. I would recognize, as the Court apparently cannot, that “the State’s knowledge of [an] individual’s predicament [and] its expressions of intent to help him” can amount to a “limitation . . . on his freedom to act on his own behalf” or to obtain help from others.

[Similarly, i]n striking down a filing fee as applied to divorce cases brought by indigents and in deciding that a local government could not entirely foreclose the opportunity to speak in a public forum, we have acknowledged that a State’s actions—such as the monopolization of a particular path of relief—may impose upon the State certain positive duties. Similarly, *Shelley* and *Burton* suggest that a State may be found complicit in an injury even if it did not create the situation that caused the harm.

. . . [Thus,] a State’s prior actions may be decisive in analyzing the constitutional significance of its inaction. I thus would locate the DeShaneys’ claims within th[at] framework . . . by considering the actions that Wisconsin took with respect to Joshua.
Wisconsin has established a child-welfare system specifically designed to help children like Joshua. Wisconsin law places upon the local departments of social services . . . a duty to investigate reported instances of child abuse. While other governmental bodies and private persons are largely responsible for the reporting of possible cases of child abuse, Wisconsin law channels all such reports to the local departments of social services for evaluation and, if necessary, further action. Even when it is the sheriff’s office or police department that receives a report of suspected child abuse, that report is referred to local social services departments for action; the only exception to this occurs when the reporter fears for the child’s immediate safety. In this way, Wisconsin law invites—indeed, directs—citizens and other governmental entities to depend on local departments of social services such as respondent to protect children from abuse.

[Moreover,] the Department[ ] controlled the decision whether to take steps to protect a particular child from suspected abuse. While many different people contributed information and advice to this decision, it was up to the people at DSS to make the ultimate decision . . . whether to disturb the family’s current arrangements. When Joshua first appeared at a local hospital with injuries signaling physical abuse, for example, it was DSS that made the decision to take him into temporary custody for the purpose of studying his situation—and it was DSS . . . that returned him to his father. Unfortunately for Joshua DeShaney, the buck effectively stopped with the Department.

In these circumstances, a private citizen, or even a person working in a government agency other than DSS, would doubtless feel that her job was done as soon as she had reported her suspicions of child abuse to DSS. Through its child-welfare program, in other words, the State of Wisconsin has relieved ordinary citizens and governmental bodies other than the Department of any sense of obligation to do anything more than report their suspicions of child abuse to DSS. If DSS ignores or dismisses these suspicions, no one will step in to fill the gap. Wisconsin’s child-protection program thus effectively confined Joshua DeShaney within the walls of [his father’s] violent home until such time as DSS took action to remove him. Conceivably, then, children like Joshua are made worse off by the existence of this program when the persons and entities charged with carrying it out fail to do their jobs.

It simply belies reality, therefore, to contend that the State “stood by and did nothing” with respect to Joshua. Through its child-protection program, the State actively intervened in Joshua’s life and, by virtue of this intervention, acquired ever more certain knowledge that Joshua was in grave danger. These circumstances, in my view, plant this case solidly within the tradition of [the incarceration and institutionalization cases] . . . .

[One may want to] defer[] to a decisionmaker’s professional judgment[, so] that once a caseworker has decided, on the basis of her professional training and experience, that one course of protection is preferable for a given child, or even that no special protection is required, she will not be found liable for the harm that follows. . . . Moreover, that the Due Process Clause is not violated by merely negligent conduct means that a social worker who simply makes a mistake of judgment under what are admittedly complex and difficult conditions will not find herself liable in damages under § 1983.
As the Court today reminds us, “the Due Process Clause of the Fourteenth Amendment was intended to prevent government ‘from abusing [its] power, or employing it as an instrument of oppression.’” My disagreement with the Court arises from its failure to see that inaction can be every bit as abusive of power as action, that oppression can result when a State undertakes a vital duty and then ignores it. Today’s opinion construes the Due Process Clause to permit a State to displace private sources of protection and then, at the critical moment, to shrug its shoulders and turn away from the harm that it has promised to try to prevent. Because I cannot agree that our Constitution is indifferent to such indifference, I respectfully dissent.

**JUSTICE BLACKMUN, dissenting.**

Today, the Court purports to be the dispassionate oracle of the law, unmoved by “natural sympathy.” But, in this pretense, the Court itself retreats into a sterile formalism which prevents it from recognizing either the facts of the case before it or the legal norms that should apply to those facts. As JUSTICE BRENNAN demonstrates, the facts here involve not mere passivity, but active state intervention in the life of Joshua DeShaney—intervention that triggered a fundamental duty to aid the boy once the State learned of the severe danger to which he was exposed.

The Court fails to recognize this duty because it attempts to draw a sharp and rigid line between action and inaction. But such formalistic reasoning has no place in the interpretation of the broad and stirring Clauses of the Fourteenth Amendment. Indeed, I submit that these Clauses were designed, at least in part, to undo the formalistic legal reasoning that infected antebellum jurisprudence, which the late Professor Robert Cover analyzed so effectively in his significant work entitled *Justice Accused* (1975).

Like the antebellum judges who denied relief to fugitive slaves, the Court today claims that its decision, however harsh, is compelled by existing legal doctrine. On the contrary, the question presented by this case is an open one, and our Fourteenth Amendment precedents may be read more broadly or narrowly depending upon how one chooses to read them. Faced with the choice, I would adopt a “sympathetic” reading, one which comports with dictates of fundamental justice and recognizes that compassion need not be exiled from the province of judging.

Poor Joshua! Victim of repeated attacks by an irresponsible, bullying, cowardly, and intemperate father, and abandoned by respondents who placed him in a dangerous predicament and who knew or learned what was going on, and yet did essentially nothing except, as the Court revealingly observes, “dutifully recorded these incidents in [their] files.” It is a sad commentary upon American life, and constitutional principles—so full of late of patriotic fervor and proud proclamations about “liberty and justice for all”—that this child, Joshua DeShaney, now is assigned to live out the remainder of his life profoundly retarded. Joshua and his mother, as petitioners here, deserve—but now are denied by this Court—the opportunity to have the facts of their case considered in the light of the constitutional protection that 42 U.S.C. § 1983 is meant to provide.
§ 2.12. Notes on the conceptual difficulties of the state-action doctrine

1. The importance of the state action doctrine. Charles Black has said that state action “is the most important problem in American law. We cannot think about it too much; we ought to take about it until we settle on a view both conceptually and functionally right.” Black, Foreword: “State Action,” Equal Protection, and California’s Proposition 14, 81 HARV. L. REV. 69, 70 (1967). Because, as we have seen, state action makes the difference between the applicability and the non-applicability of many important constitutional provisions—and because expansive readings of state action, as in Shelley, could make all of private life subject to constitutional restrictions—this is perhaps not an exaggeration.

2. The “incoherence” of the state action doctrine. Black called state action a “conceptual disaster area” in 1967. Fifteen years later (and half a year before the Rendell-Baker/Blum/Lugar trilogy, Judge Friendly said that characterization was “even more apt today.” Even the Supreme Court, in Lebron v. National Railroad Passenger Corp. (1995), admitted that its cases “have not been a model of consistency.” But see Tribe, CONSTITUTIONAL CHOICES 248 (1985) (saying there’s more consistency than there seems to be). Chemerinsky writes that the generation of commentators from the 1940s to the 1970s “successful[ly] demonstrat[ed] the incoherence” of the doctrine and “persuasively argued that the concept of state action never could be rationally or consistently applied.” Chemerinsky, Rethinking State Action, 80 NW. L. REV. 503 (1985).

3. The persistence of the state action doctrine. Jerre Williams wrote in 1963 not just that the state action doctrine was incoherent, but also that it was dying out in the cases. “The sun is setting on the concept of state action”—after cases like Shelley and Burton, everything is state action, and we can proceed directly to an analysis on the merits of whether the challenged behavior was unconstitutional. Thus, private discriminatory behavior would be subject to the Equal Protection Clause, but could be upheld sometimes based on competing concerns, for instance freedom of private association for certain organizations, or the affirmative value of religious discrimination for parochial schools. Williams, The Twilight of State Action, 41 TEX. L. REV. 347 (1963).

Nonetheless, says Chemerinsky, the Burger Court showed no interest in rolling the doctrine back, and, indeed, reinvigorated it. As Louis Michael Seidman writes, “we keep obsessively returning to the state action problem long after the basic analytics have been explicated and well understood.” Seidman, The State Action Paradox, 10 CONST. COMM. 379 (1993).

Even if one believes the state action doctrine to be incoherent, would an open-ended balancing approach be a good idea? Should all issues of everyday life become federal constitutional issues? See also note 4 infra. Can judges be trusted to strike the right balance? (On the perils of balancing, cf. Eugene Volokh,
Calls for rolling back the state action limitation (that is, for expanding the category of things subject to constitutional constraints) usually come from the political left; but, unless the judges are also on the political left, could the result be worse from the critics’ perspective?

4. **Seidman on the tensions in state action doctrine.** Seidman explains that the problematic state action doctrine “is a symptom of a larger problem. The problem is produced by the partial but incomplete assimilation into our legal culture of realist insights from half a century ago. In the modern period, we can neither escape nor fully accept these insights, and our ambivalent reaction to them has created antinomies that obstruct not only the creation of a coherent view of state action, but also the possibility of serious constitutional discourse on almost any subject.”

One of the central goals of the Legal Realist movement was to dismantle the distinction between public and private, between action and inaction—to see the action of government everywhere, even in its non-intervention, and to see the rights formerly thought of as “natural” as resting on entitlements created by government enforcement. This move paved the way for an argument that government reallocation of property rights—as in the New Deal—was merely a question of policy, not (as the *Lochner*-era judges believed) a question of right. But New Deal-era judges could not fully endorse this entire reasoning, for if reallocation of rights is merely a political question, what is the role of unelected judges? Indeed, this was their very critique of the *Lochner* judges—that they were exercising a political power disguised as neutral judging. Was every policy question now to become a question of constitutional law to be settled, again, by unelected judges? Therefore, the very judges that criticized the state action doctrine were forced to resurrect it in order to protect political questions from judicial resolution.

Moreover, the idea that certain rights are individual and cannot be attributed to the state—whether the *Lochner*-era idea that we have a right to choose the education of our children, or the modern liberal idea that we have a right to choose our sexuality or abortions—is deeply ingrained into our modern culture. Therefore, the critique of the state action doctrine runs up against deeply held beliefs in a certain sphere of autonomy.

5. **The tension as applied to *DeShaney*.** Thus, in *DeShaney*, Seidman argues, Brennan is right to point out that Rehnquist’s action/inaction holding is unsatisfactory. One might ask why it makes sense to subject an activist social worker, who removes Joshua without authorization, to constitutional requirements, but completely immunize the evil social worker who stands by and does nothing because he wants to see Joshua die.

However, Brennan does not pursue this path—in fact, he does not criticize the state action doctrine, or the action/inaction distinction, at all. (Brennan apparently
agrees that the state would not be liable if it abandoned the welfare business entirely.) Instead, he urges us to see action where Rehnquist sees inaction; inaction is always embedded in a network of actions, and the Wisconsin government did engage in a great many “actions” in this case, some of which might have made Joshua worse off. This argument is highly speculative (is a do-nothing DSS really worse for Joshua than no DSS at all?), and ignores one obvious state action: the state’s decision to grant custody of Joshua to his father. However, to talk of family courts’ custody decisions as allocating infants to their biological parents is false to our deep intuitions about childrearing; parents naturally have custody of their children; the state does not grant it to them.

Meanwhile, Blackmun’s dissent, while cryptic, seems to pave the way for a more wide-ranging analysis, abandoning formalism, informed by justice and compassion, and perhaps imposing more affirmative duties on state officials. But since no court can undo the damage suffered by Joshua, any such proposal must stand or fall based on its effects in future cases. Would saving more Joshuas counterbalance the intrusion into the family life of parents who end up being found innocent? Would allocating more funds to the DSS be worth the corresponding reductions in other services? This is a difficult question, which compassion alone cannot answer. And this critique of the action/inaction distinction is doubly problematic, coming as it does from the author of Roe v. Wade, which stands for the idea that certain choices are fundamentally private.


CHIEF JUSTICE REHNQUIST delivered the opinion of the Court.

Pennsylvania provides in its workers’ compensation regime that an employer or insurer may withhold payment for disputed medical treatment pending an independent review to determine whether the treatment is reasonable and necessary. We hold that the insurers are not “state actors” under the Fourteenth Amendment . . .

I

. . . Pennsylvania’s Workers’ Compensation Act . . . creates a system of no-fault liability for work-related injuries and makes employers’ liability under this system “exclusive . . . of any and all other liability.” All employers subject to the Act must (1) obtain workers’ compensation insurance from a private insurer, (2) obtain such insurance through the State Workmen’s Insurance Fund (SWIF), or (3) seek permission from the State to self-insure. Once an employer becomes liable for an employee’s work-related injury—because liability either is not contested or is no longer at issue—the employer or its insurer must pay for all “reasonable” and “necessary” medical treatment, and must do so within 30 days of receiving a bill.
To assure that insurers pay only for medical care that meets these criteria, and in an attempt to control costs, [later amendments to the statute] created a “utilization review” procedure under which the reasonableness and necessity of an employee’s . . . medical treatment could be reviewed before a medical bill must be paid. . . . If an insurer “disputes the reasonableness or necessity of the treatment provided,” it may request utilization review (within the same 30-day period) by filing a one-page form with the Workers’ Compensation Bureau of the Pennsylvania Department of Labor and Industry (Bureau). The form identifies (among other things) the employee, the medical provider, the date of the employee’s injury, and the medical treatment to be reviewed. . . . Upon the filing of a request, an insurer may withhold payment to health care providers for the particular services being challenged.

The Bureau then notifies the parties that utilization review has been requested and forwards the request to a randomly selected “utilization review organization” (URO). URO’s are private organizations consisting of health care providers who are “licensed in the same profession and hav[en] the same or similar specialty as that of the provider of the treatment under review.” The purpose of utilization review, and the sole authority conferred upon a URO, is to determine “whether the treatment under review is reasonable or necessary for the medical condition of the employee” in light of “generally accepted treatment protocols.” Reviewers must examine the treating provider’s medical records, and must give the provider an opportunity to discuss the treatment under review. Any doubt as to the reasonableness and necessity of a given procedure must be resolved in favor of the employee.

URO’s are instructed to complete their review and render a determination within 30 days of a completed request. If the URO finds in favor of the insurer, the employee may appeal the determination to a workers’ compensation judge for a de novo review, but the insurer need not pay for the disputed services unless the URO’s determination is overturned by the judge, or later by the courts. If the URO finds in favor of the employee, the insurer must pay the disputed bill immediately, with 10 percent annual interest, as well as the cost of the utilization review.

[Patients who had] payment of particular benefits withheld pursuant to the utilization review procedure . . . sued [some government actors and private insurers] under 42 U.S.C. § 1983, . . . alleg[ing] that in withholding . . . benefits without predeprivation notice and an opportunity to be heard, . . . defendants, acting “under color of state law,” deprived them of property in violation of due process. . . .

II

To state a claim . . . under § 1983, respondents must establish that they were deprived of a right secured by the Constitution or laws of the United States, and that the alleged deprivation was committed under color of state law. Like the state-action requirement of
the Fourteenth Amendment, the under-color-of-state-law element of § 1983 excludes from its reach “‘merely private conduct, no matter how discriminatory or wrongful.’”8

Perhaps hoping to avoid the traditional application of our state-action cases, respondents attempt to characterize their claim as a “facial” or “direct” challenge to the utilization review procedures contained in the Act, in which case, the argument goes, we need not concern ourselves with the “identity of the defendant” or the “act or decision by a private actor or entity who is relying on the challenged law.” This argument, however, ignores our repeated insistence that state action requires both an alleged constitutional deprivation “caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the State or by a person for whom the State is responsible,” and that “the party charged with the deprivation must be a person who may fairly be said to be a state actor.” See Lugar. In this case, while it may fairly be said that private insurers act “‘with the knowledge of and pursuant to’” the state statute, thus satisfying the first requirement, respondents still must satisfy the second, whether the allegedly unconstitutional conduct is fairly attributable to the State.

Our approach to this latter question begins by identifying “the specific conduct of which the plaintiff complains.” Here, respondents named as defendants both public officials and a class of private insurers and self-insured employers. . . . The complaint alleged that the state and private defendants, acting under color of state law and pursuant to the Act, deprived them of property in violation of due process by withholding payment for medical treatment without prior notice and an opportunity to be heard. All agree that the public officials responsible for administering the workers’ compensation system and the director of SWIF are state actors. Thus, the issue we address, in accordance with our cases, is whether a private insurer’s decision to withhold payment for disputed medical treatment may be fairly attributable to the State so as to subject insurers to the constraints of the Fourteenth Amendment. Our answer to that question is “no.”

In cases involving extensive state regulation of private activity, we have consistently held that “[t]he mere fact that a business is subject to state regulation does not by itself convert its action into that of the State for purposes of the Fourteenth Amendment.” Faithful application of the state-action requirement in these cases ensures that the prerogative of regulating private business remains with the States and the representative branches, not the courts. Thus, the private insurers in this case will not be held to constitutional standards unless “there is a sufficiently close nexus between the State and the challenged action of the regulated entity so that the action of the latter may be fairly treated as that of the State itself.” Whether such a “close nexus” exists, our cases state, depends on whether the State “has exercised coercive power or has provided such significant encouragement, either overt or covert, that the choice must in law be deemed to be that of the State.” Action taken by private entities with the mere approval or acquiescence of the State is not state action.

8 Where, as here, deprivations of rights under the Fourteenth Amendment are alleged, these two requirements converge. See Lugar.
Here, respondents do not assert that the decision to invoke utilization review should be attributed to the State because the State compels or is directly involved in that decision. Obviously the State is not so involved. It authorizes, but does not require, insurers to withhold payments for disputed medical treatment. The decision to withhold payment, like the decision to transfer Medicaid patients to a lower level of care in *Blum*, is made by concededly private parties, and “turns on . . . judgments made by private parties” without “standards . . . established by the State.”

Respondents do assert, however, that the decision to withhold payment to providers may be fairly attributable to the State because the State has “authorized” and “encouraged” it. Respondents’ primary argument in this regard is that, in amending the Act to provide for utilization review and to grant insurers an option they previously did not have, the State purposely “encouraged” insurers to withhold payments for disputed medical treatment. This argument reads too much into the State’s reform, and in any event cannot be squared with our cases.

We do not doubt that the State’s decision to provide insurers the option of deferring payment for unnecessary and unreasonable treatment pending review can in some sense be seen as encouraging them to do just that. But, as petitioners note, this kind of subtle encouragement is no more significant than that which inheres in the State’s creation or modification of any legal remedy. We have never held that the mere availability of a remedy for wrongful conduct, even when the private use of that remedy serves important public interests, so significantly encourages the private activity as to make the State responsible for it. It bears repeating that a finding of state action on this basis would be contrary to the “essential dichotomy” between public and private acts that our cases have consistently recognized.

The State’s decision to allow insurers to withhold payments pending review can just as easily be seen as state inaction, or more accurately, a legislative decision not to intervene in a dispute between an insurer and an employee over whether a particular treatment is reasonable and necessary. Before the . . . amendments, Pennsylvania restricted the ability of an insurer (after liability had been established, of course) to defer workers’ compensation medical benefits, including payment for unreasonable and unnecessary treatment, beyond 30 days of receipt of the bill. The . . . amendments, in effect, restored to insurers the narrow option, historically exercised by employers and insurers before the adoption of Pennsylvania’s workers’ compensation law, to defer payment of a bill until it is substantiated. The most that can be said of the statutory scheme, therefore, is that whereas it previously prohibited insurers from withholding payment for disputed medical services, it no longer does so. Such permission of a private choice cannot support a finding of state action. As we have said before, our cases will not tolerate “the imposition of Fourteenth Amendment restraints on private action by the simple device of characterizing the State’s inaction as ‘authorization’ or ‘encouragement.’”

Nor does the State’s role in creating, supervising, and setting standards for the URO process differ in any meaningful sense from the creation and administration of any forum for resolving disputes. While the decision of a URO, like that of any judicial official, may
properly be considered state action, a private party’s mere use of the State’s dispute resolution machinery, without the “overt, significant assistance of state officials,” cannot.

The State, in the course of administering a many-faceted remedial system, has shifted one facet from favoring the employees to favoring the employer. This sort of decision occurs regularly in legislative review of such systems. But it cannot be said that such a change “encourages” or “authorizes” the insurer’s actions as those terms are used in our state-action jurisprudence.

We also reject the notion . . . that the challenged decisions are state action because insurers must first obtain “authorization” or “permission” from the Bureau before withholding payment. . . . [T]he Bureau’s participation is limited to requiring insurers to file “a form prescribed by the Bureau,” processing the request for technical compliance, and then forwarding the matter to a URO and informing the parties that utilization review has been requested. In Blum, we rejected the notion that the State, “by requiring completion of a form,” is responsible for the private party’s decision. The additional “paper shuffling” performed by the Bureau here in response to an insurers’ request does not alter that conclusion.

Respondents next contend that state action is present because the State has delegated to insurers “powers traditionally exclusively reserved to the State.” Their argument here is twofold. Relying on West, respondents first argue that workers’ compensation benefits are state-mandated “public benefits,” and that the State has delegated the provision of these “public benefits” to private insurers. They also contend that the State has delegated to insurers the traditionally exclusive government function of determining whether and under what circumstances an injured worker’s medical benefits may be suspended. . . .

We think neither argument has merit. West is readily distinguishable: There the State was constitutionally obligated to provide medical treatment to injured inmates, and the delegation of that traditionally exclusive public function to a private physician gave rise to a finding of state action. Here, on the other hand, nothing in Pennsylvania’s Constitution or statutory scheme obligates the State to provide either medical treatment or workers’ compensation benefits to injured workers. Instead, the State’s workers’ compensation law imposes that obligation on employers. This case is therefore not unlike Jackson v. Metropolitan Edison Co. (1974), where we noted that “while the Pennsylvania statute imposes an obligation to furnish service on regulated utilities, it imposes no such obligation on the State.”

Nor is there any merit in respondents’ argument that the State has delegated to insurers the traditionally exclusive governmental function of deciding whether to suspend payment for disputed medical treatment. Historical practice, as well as the state statutory scheme, does not support respondents’ characterization. It is no doubt true that before the . . . amendments an insurer who sought to withhold payment for disputed medical treatment was required to petition the Bureau, and could withhold payment only upon a favorable ruling by a workers’ compensation judge, and then only for prospective treatment.
But before Pennsylvania ever adopted its workers’ compensation law, an insurer under contract with an employer to pay for its workers’ reasonable and necessary medical expenses could withhold payment, for any reason or no reason, without any authorization or involvement of the State. The insurer, of course, might become liable to the employer (or its workers) if the refusal to pay breached the contract or constituted “bad faith,” but the obligation to pay would only arise after the employer had initiated a claim and reduced it to a judgment. That Pennsylvania first recognized an insurer’s traditionally private prerogative to withhold payment, then restricted it, and now (in one limited respect) has restored it, cannot constitute the delegation of a traditionally exclusive public function. Like New York in *Flagg Bros.*, Pennsylvania “has done nothing more than authorize (and indeed limit)—without participation by any public official—what [private insurers] would tend to do, even in the absence of such authorization,” i.e., withhold payment for disputed medical treatment pending a determination that the treatment is, in fact, reasonable and necessary. . . .

. . . [R]espondents [also] urge us to [adopt the] “joint participation” theory of state action [of *Burton* and hold that the Pennsylvania system “inextricably entangles the insurance companies in a partnership with the Commonwealth such that they become an integral part of the state in administering the statutory scheme.”]

*Burton* was one of our early cases dealing with “state action” under the Fourteenth Amendment, and later cases have refined the vague “joint participation” test embodied in that case. *Blum* and *Jackson*, in particular, have established that “privately owned enterprises providing services that the State would not necessarily provide, even though they are extensively regulated, do not fall within the ambit of *Burton*.” Here, workers’ compensation insurers are at least as extensively regulated as the private nursing facilities in *Blum* and the private utility in *Jackson*. Like those cases, though, the state statutory and regulatory scheme leaves the challenged decisions to the judgment of insurers.

Respondents also rely on *Lugar*, which contains general language about “joint participation” as a test for state action. But, as the *Lugar* opinion itself makes clear, its language must not be torn from the context out of which it arose: “. . . a system whereby state officials will attach property on the ex parte application of one party to a private dispute.” . . .

We conclude that an insurer’s decision to withhold payment and seek utilization review of the reasonableness and necessity of particular medical treatment is not fairly attributable to the State. . . .

**JUSTICE STEVENS, concurring in part and dissenting in part.**

. . . [T]he Court[’s] . . . opinion . . . incorrectly assumes that the question whether the insurance company is a state actor is relevant to the controlling question whether the state procedures are fair. The relevant state actors, rather than the particular parties to the payment disputes, are the state-appointed decisionmakers who implement the exclusive procedure that the State has created to protect respondents’ rights. These state actors are defendants in this suit. . . .

JUSTICE SOUTER delivered the opinion of the Court.

The issue is whether a statewide association incorporated to regulate interscholastic athletic competition among public and private secondary schools may be regarded as engaging in state action when it enforces a rule against a member school. The association in question here includes most public schools located within the State, acts through their representatives, draws its officers from them, is largely funded by their dues and income received in their stead, and has historically been seen to regulate in lieu of the State Board of Education’s exercise of its own authority. We hold that the association’s regulatory activity may and should be treated as state action owing to the pervasive entwinement of state school officials in the structure of the association, there being no offsetting reason to see the association’s acts in any other way.

I

Respondent Tennessee Secondary School Athletic Association (Association) is a not-for-profit membership corporation organized to regulate interscholastic sport among the public and private high schools in Tennessee that belong to it. No school is forced to join, but without any other authority actually regulating interscholastic athletics, it enjoys the memberships of almost all the State’s public high schools (some 290 of them or 84% of the Association’s voting membership), far outnumbering the 55 private schools that belong. A member school’s team may play or scrimmage only against the team of another member, absent a dispensation.

The Association’s rulemaking arm is its legislative council, while its board of control tends to administration. The voting membership of each of these nine-person committees is limited under the Association’s bylaws to high school principals, assistant principals, and superintendents elected by the member schools, and the public school administrators who so serve typically attend meetings during regular school hours. Although the Association’s staff members are not paid by the State, they are eligible to join the State’s public retirement system for its employees. Member schools pay dues to the Association, though the bulk of its revenue is gate receipts at member teams’ football and basketball tournaments, many of them held in public arenas rented by the Association.

The constitution, bylaws, and rules of the Association set standards of school membership and the eligibility of students to play in interscholastic games. Each school, for example, is regulated in awarding financial aid, most coaches must have a Tennessee state teaching license, and players must meet minimum academic standards and hew to limits on student employment. Under the bylaws, “in all matters pertaining to the athletic relations of his school,” the principal is responsible to the Association, which has the power “to suspend, to fine, or otherwise penalize any member school for the violation of any of the rules of the Association or for other just cause.”

Draft — do not distribute!
Ever since the Association was incorporated in 1925, Tennessee’s State Board of Education (State Board) has (to use its own words) acknowledged the corporation’s functions “in providing standards, rules and regulations for interscholastic competition in the public schools of Tennessee.” More recently, the State Board cited its statutory authority when it adopted language expressing the relationship between the Association and the State Board. Specifically, in 1972, it went so far as to adopt a rule expressly “designating” the Association as “the organization to supervise and regulate the athletic activities in which the public junior and senior high schools in Tennessee participate on an interscholastic basis.” The Rule provided that “the authority granted herein shall remain in effect until revoked” and instructed the State Board’s chairman to “designate a person or persons to serve in an ex-officio capacity on the [Association’s governing bodies].” That same year, the State Board specifically approved the Association’s rules and regulations, while reserving the right to review future changes. Thus, on several occasions over the next 20 years, the State Board reviewed, approved, or reaffirmed its approval of the recruiting Rule at issue in this case. In 1996, however, the State Board dropped the original Rule . . . expressly designating the Association as regulator; it substituted a statement “recognizing the value of participation in interscholastic athletics and the role of [the Association] in coordinating interscholastic athletic competition,” while “authorizing the public schools of the state to voluntarily maintain membership in [the Association].”

The action before us responds to a 1997 regulatory enforcement proceeding brought against petitioner, Brentwood Academy, a private parochial high school member of the Association. The Association’s board of control found that Brentwood violated a rule prohibiting “undue influence” in recruiting athletes, when it wrote to incoming students and their parents about spring football practice. The Association accordingly placed Brentwood’s athletic program on probation for four years, declared its football and boys’ basketball teams ineligible to compete in playoffs for two years, and imposed a $3,000 fine. When these penalties were imposed, all the voting members of the board of control and legislative council were public school administrators.

Brentwood sued the Association and its executive director in federal court under . . . 42 U.S.C. § 1983, claiming that enforcement of the Rule was state action and a violation of the First and Fourteenth Amendments. . . .

II

A

Our cases try to plot a line between state action subject to Fourteenth Amendment scrutiny and private conduct (however exceptionable) that is not. The judicial obligation is not only to “‘preserve[e] an area of individual freedom by limiting the reach of federal law’ and avoid[d] the imposition of responsibility on a State for conduct it could not control,” but also to assure that constitutional standards are invoked “when it can be said that the State is responsible for the specific conduct of which the plaintiff complains.” If the Fourteenth Amendment is not to be displaced, therefore, its ambit cannot be a simple line between States and people operating outside formally governmental organizations,
and the deed of an ostensibly private organization or individual is to be treated sometimes as if a State had caused it to be performed. Thus, we say that state action may be found if, though only if, there is such a “close nexus between the State and the challenged action” that seemingly private behavior “may be fairly treated as that of the State itself.”

What is fairly attributable is a matter of normative judgment, and the criteria lack rigid simplicity. From the range of circumstances that could point toward the State behind an individual face, no one fact can function as a necessary condition across the board for finding state action; nor is any set of circumstances absolutely sufficient, for there may be some countervailing reason against attributing activity to the government.

Our cases have identified a host of facts that can bear on the fairness of such an attribution. We have, for example, held that a challenged activity may be state action when it results from the State’s exercise of “coercive power,” when the State provides “significant encouragement, either overt or covert,” or when a private actor operates as a “willful participant in joint activity with the State or its agents.” We have treated a nominally private entity as a state actor when it is controlled by an “agency of the State,” when it has been delegated a public function by the State, when it is “entwined with governmental policies,” or when government is “entwined in [its] management or control.”

Amidst such variety, examples may be the best teachers, and examples from our cases are unequivocal in showing that the character of a legal entity is determined neither by its expressly private characterization in statutory law, nor by the failure of the law to acknowledge the entity’s inseparability from recognized government officials or agencies. . . . Amtrak [is] the Government for constitutional purposes, regardless of its congressional designation as private; it was organized under federal law to attain governmental objectives and was directed and controlled by federal appointees. . . . [T]he privately endowed Girard College [was held] to be a state actor and enforcement of its private founder’s limitation of admission to whites [was held to be] attributable to the State, because, consistent with the terms of the settlor’s gift, the college’s board of directors was a state agency established by state law. [Conversely,] private trustees to whom a city had transferred a park were nonetheless [held to be] state actors barred from enforcing racial segregation, since the park served the public purpose of providing community recreation, and “the municipality remain[ed] entwined in [its] management [and] control.”

These examples of public entwinement in the management and control of ostensibly separate trusts or corporations foreshadow this case, as this Court itself anticipated in National Collegiate Athletic Ass’n v. Tarkanian. Tarkanian arose when an undoubtedly state actor, the University of Nevada, suspended its basketball coach, Tarkanian, in order to comply with rules and recommendations of the National Collegiate Athletic Association (NCAA). The coach charged the NCAA with state action, arguing that the state university had delegated its own functions to the NCAA, clothing the latter with authority to make and apply the university’s rules, the result being joint action making the NCAA a state actor.
To be sure, it is not the strict holding in *Tarkanian* that points to our view of this case, for we found no state action on the part of the NCAA. We could see, on the one hand, that the university had some part in setting the NCAA’s rules, and the Supreme Court of Nevada had gone so far as to hold that the NCAA had been delegated the university’s traditionally exclusive public authority over personnel. But on the other side, the NCAA’s policies were shaped not by the University of Nevada alone, but by several hundred member institutions, most of them having no connection with Nevada, and exhibiting no color of Nevada law. Since it was difficult to see the NCAA, not as a collective membership, but as surrogate for the one State, we held the organization’s connection with Nevada too insubstantial to ground a state-action claim.

But dictum in *Tarkanian* pointed to a contrary result on facts like ours, with an organization whose member public schools are all within a single State. “The situation would, of course, be different if the [Association’s] membership consisted entirely of institutions located within the same State, many of them public institutions created by the same sovereign.” . . .

B

Just as we foresaw in *Tarkanian*, the “necessarily fact-bound inquiry” leads to the conclusion of state action here. The nominally private character of the Association is overborne by the pervasive entwinement of public institutions and public officials in its composition and workings, and there is no substantial reason to claim unfairness in applying constitutional standards to it.

The Association is not an organization of natural persons acting on their own, but of schools, and of public schools to the extent of 84% of the total. Under the Association’s bylaws, each member school is represented by its principal or a faculty member, who has a vote in selecting members of the governing legislative council and board of control from eligible principals, assistant principals, and superintendents.

Although the findings and prior opinions in this case include no express conclusion of law that public school officials act within the scope of their duties when they represent their institutions, no other view would be rational, the official nature of their involvement being shown in any number of ways. Interscholastic athletics obviously play an integral part in the public education of Tennessee, where nearly every public high school spends money on competitions among schools. Since a pickup system of interscholastic games would not do, these public teams need some mechanism to produce rules and regulate competition. The mechanism is an organization overwhelmingly composed of public school officials who select representatives (all of them public officials at the time in question here), who in turn adopt and enforce the rules that make the system work. Thus, by giving these jobs to the Association, the 290 public schools of Tennessee belonging to it can sensibly be seen as exercising their own authority to meet their own responsibilities. Unsurprisingly, then, the record indicates that half the council or board meetings documented here were held during official school hours, and that public schools have largely provided for the Association’s financial support. A small portion of the Association’s revenue comes from membership dues paid by the schools, and the

Draft — do not distribute!
principal part from gate receipts at tournaments among the member schools. Unlike mere public buyers of contract services, whose payments for services rendered do not convert the service providers into public actors, the schools here obtain membership in the service organization and give up sources of their own income to their collective association. The Association thus exercises the authority of the predominantly public schools to charge for admission to their games; the Association does not receive this money from the schools, but enjoys the schools’ moneymaking capacity as its own.

In sum, to the extent of 84% of its membership, the Association is an organization of public schools represented by their officials acting in their official capacity to provide an integral element of secondary public schooling. There would be no recognizable Association, legal or tangible, without the public school officials, who do not merely control but overwhelmingly perform all but the purely ministerial acts by which the Association exists and functions in practical terms. Only the 16% minority of private school memberships prevents this entwinement of the Association and the public school system from being total and their identities totally indistinguishable.

To complement the entwinement of public school officials with the Association from the bottom up, the State of Tennessee has provided for entwinement from top down. State Board members are assigned ex officio to serve as members of the board of control and legislative council, and the Association’s ministerial employees are treated as state employees to the extent of being eligible for membership in the state retirement system.

It is, of course, true that the time is long past when the close relationship between the surrogate association and its public members and public officials acting as such was attested frankly. As mentioned, the terms of the State Board’s Rule expressly designating the Association as regulator of interscholastic athletics in public schools were deleted in 1996, the year after a Federal District Court held that the Association was a state actor because its rules were “caused, directed and controlled by the Tennessee Board of Education.”

But the removal of the designation language from [the] Rule . . . affected nothing but words. Today the State Board’s member-designees continue to sit on the Association’s committees as nonvoting members, and the State continues to welcome Association employees in its retirement scheme. The close relationship is confirmed by the Association’s enforcement of the same preamendment rules and regulations reviewed and approved by the State Board (including the recruiting Rule challenged by Brentwood), and by the State Board’s continued willingness to allow students to satisfy its physical education requirement by taking part in interscholastic athletics sponsored by the Association. The most one can say on the evidence is that the State Board once freely acknowledged the Association’s official character but now does it by winks and nods.

---

3 That District Court . . . held that “[t]his delegation of authority to TSSAA to Tennessee, standing alone, is sufficient to make TSSAA a state actor” under the “state compulsion test,” which it understood to provide that a State could exercise such coercive power or provide such significant encouragement, either overt or covert, that the choice of the private actor must be deemed to be that of the State as a matter of law.

4 The significance of winks and nods in state-action doctrine seems to be one of the points of the dissenters’ departure from the rest of the Court. In drawing the public-private action line, the dissenters would
The amendment to the Rule in 1996 affected candor but not the “momentum” of the Association’s prior involvement with the State Board. . . . [Because of] “custom and practice,” “the conduct of the parties has not materially changed” since 1996, “the connections between TSSAA and the State [being] still pervasive and entwined.”

The entwinement down from the State Board is therefore unmistakable, just as the entwinement up from the member public schools is overwhelming. Entwinement will support a conclusion that an ostensibly private organization ought to be charged with a public character and judged by constitutional standards; entwinement to the degree shown here requires it. . . .

D

. . . [However, e]ven facts that suffice to show public action (or, standing alone, would require such a finding) may be outweighed in the name of some value at odds with finding public accountability in the circumstances. In Polk County, a defense lawyer’s actions were deemed private even though she was employed by the county and was acting within the scope of her duty as a public defender. Full-time public employment would be conclusive of state action for some purposes, see West, but not when the employee is doing a defense lawyer’s primary job; then, the public defender does “not act on behalf of the State; he is the State’s adversary.” The state-action doctrine does not convert opponents into virtual agents.

The assertion of such a countervailing value is the nub of each of the Association’s two remaining arguments, neither of which, however, persuades us. The Association suggests, first, that reversing the judgment here will somehow trigger an epidemic of unprecedented federal litigation. Even if that might be counted as a good reason for a Polk County decision to call the Association’s action private, [there is no evidence that this would happen] . . . .

Nor do we think there is anything to be said for the Association’s contention that there is no need to treat it as a state actor since any public school applying the Association’s rules is itself subject to suit under § 1983 or Title IX of the Education Amendments of 1972. If Brentwood’s claim were pushing at the edge of the class of possible defendant state actors, an argument about the social utility of expanding that class would at least be on point, but because we are nowhere near the margin in this case, the Association is really asking for nothing less than a dispensation for itself. Its position boils down to saying that the Association should not be dressed in state clothes because other, concededly public actors are; that Brentwood should be kept out of court because a different plaintiff raising
a different claim in a different case may find the courthouse open. Pleas for special
treatment are hard to sell, although saying that does not, of course, imply anything about
the merits of Brentwood’s complaint; the issue here is merely whether Brentwood
properly names the Association as a § 1983 defendant, not whether it should win on its
claim. . . .

JUSTICE THOMAS, with whom THE CHIEF JUSTICE, JUSTICE SCALIA, and
JUSTICE KENNEDY join, dissenting.

We have never found state action based upon mere “entwinement.” Until today, we have
found a private organization’s acts to constitute state action only when the organization
performed a public function; was created, coerced, or encouraged by the government; or
acted in a symbiotic relationship with the government. The majority’s holding—that the
Tennessee Secondary School Athletic Association’s (TSSAA) enforcement of its
recruiting rule is state action—not only extends state-action doctrine beyond its
permissible limits but also encroaches upon the realm of individual freedom that the
doctrine was meant to protect. I respectfully dissent.

I

Like the state-action requirement of the Fourteenth Amendment, the state-action element
of 42 U.S.C. § 1983 excludes from its coverage “merely private conduct, however
discriminatory or wrongful.” “Careful adherence to the ‘state action’ requirement” thus
“preserves an area of individual freedom by limiting the reach of federal law and federal
judicial power.” The state-action doctrine also promotes important values of federalism,
“avoid[ing] the imposition of responsibility on a State for conduct it could not control.”
Although we have used many different tests to identify state action, they all have a
common purpose. Our goal in every case is to determine whether an action “can fairly be
attributed to the State.”

A

Regardless of these various tests for state action, common sense dictates that the
TSSAA’s actions cannot fairly be attributed to the State, and thus cannot constitute state
action. The TSSAA was formed in 1925 as a private corporation to organize
interscholastic athletics and to sponsor tournaments among its member schools. Any
private or public secondary school may join the TSSAA by signing a contract agreeing to
comply with its rules and decisions. Although public schools currently compose 84% of
the TSSAA’s membership, the TSSAA does not require that public schools constitute a
set percentage of its membership, and, indeed, no public school need join the TSSAA.
The TSSAA’s rules are enforced not by a state agency but by its own board of control,
which comprises high school principals, assistant principals, and superintendents, none of
whom must work at a public school. Of course, at the time the recruiting rule was
enforced in this case, all of the board members happened to be public school officials.
However, each board member acts in a representative capacity on behalf of all the private
and public schools in his region of Tennessee, and not simply his individual school.
The State of Tennessee did not create the TSSAA. The State does not fund the TSSAA and does not pay its employees. In fact, only 4% of the TSSAA’s revenue comes from the dues paid by member schools; the bulk of its operating budget is derived from gate receipts at tournaments it sponsors. The State does not permit the TSSAA to use state-owned facilities for a discounted fee, and it does not exempt the TSSAA from state taxation. No Tennessee law authorizes the State to coordinate interscholastic athletics or empowers another entity to organize interscholastic athletics on behalf of the State. The only state pronouncement acknowledging the TSSAA’s existence is a rule providing that the State Board of Education permits public schools to maintain membership in the TSSAA if they so choose.

Moreover, the State of Tennessee has never had any involvement in the particular action taken by the TSSAA in this case: the enforcement of the TSSAA’s recruiting rule prohibiting members from using “undue influence” on students or their parents or guardians “to secure or to retain a student for athletic purposes.” There is no indication that the State has ever had any interest in how schools choose to regulate recruiting. In fact, the TSSAA’s authority to enforce its recruiting rule arises solely from the voluntary membership contract that each member school signs, agreeing to conduct its athletics in accordance with the rules and decisions of the TSSAA.

B

[Under] the Court’s specific state-action tests, the conclusion is the same . . . .

---

1 Although the TSSAA’s employees, who typically are retired teachers, are allowed to participate in the state retirement system, the State does not pay any portion of the employer contribution for them. The TSSAA is one of three private associations, along with the Tennessee Education Association and the Tennessee School Boards Association, whose employees are statutorily permitted to participate in the state retirement system.

2 The first formal state acknowledgment of the TSSAA’s existence did not occur until 1972, when the State Board of Education passed a resolution stating that it “recognizes and designates [the TSSAA] as the organization to supervise and regulate the athletic activities in which the public junior and senior high schools of Tennessee participate in on an interscholastic basis.” There is no indication that the TSSAA invited this resolution or that the resolution in any way altered the actions of the TSSAA or the State following its adoption in 1972. In fact, it appears that the resolution was not entirely accurate: The TSSAA does not supervise or regulate regular season interscholastic contests. In any event, the resolution was revoked in 1996. Contrary to the majority’s reference to its revocation as being “winks and nods,” the repeal of the 1972 resolution appears to have had no more impact on the TSSAA’s operation than did its passage.

The majority also cites this resolution to support its assertion that “[e]ver since the Association was incorporated in 1925, Tennessee’s State Board of Education . . . has acknowledged the corporation’s function ‘in providing standards, rules and regulations for interscholastic competition in the public schools of Tennessee.’” However, there is no evidence in the record that suggests that the State of Tennessee or the State Board of Education had any involvement or interest in the TSSAA prior to 1972.

3 The Rule provides: “The State Board of Education recognizes the value of participation in interscholastic athletics and the role of the Tennessee Secondary School Athletic Association in coordinating interscholastic athletic competition. The State Board of Education authorizes the public schools of the state to voluntarily maintain membership in the Tennessee Secondary School Athletic Association.”

4 The majority relies on the fact that the TSSAA permits members of the State Board of Education to serve ex officio on its board of control to support its “top-down” theory of state action. But these members are not voting members of the TSSAA’s board of control and thus cannot exert any control over its actions.
The TSSAA has not performed a function that has been “traditionally exclusively reserved to the State.” . . . Tennessee . . . did not even show an interest in interscholastic athletics until 47 years after the TSSAA [was founded]. Even then, the State Board of Education merely acquiesced in the TSSAA’s actions . . . .

. . . [T]he TSSAA is not an entity created and controlled by the government for the purpose of fulfilling a government objective, as [is] Amtrak . . . . The TSSAA was designed to fulfill an objective . . . that the government had not contemplated, much less pursued. And although the board of control [and membership are dominated by public schools], this is not required by the TSSAA’s constitution.

. . . [T]he State of Tennessee has not “exercised coercive power or . . . provided such significant encouragement [to the TSSAA], either overt or covert,” that the TSSAA’s regulatory activities must in law be deemed to be those of the State. The State has not promulgated any regulations of interscholastic sports, and nothing in the record suggests that the State has encouraged or coerced the TSSAA in enforcing its recruiting rule. To be sure, public schools do provide a small portion of the TSSAA’s funding through their membership dues, but no one argues that these dues are somehow conditioned on the TSSAA’s enactment and enforcement of recruiting rules.5 . . . Furthermore, there is no evidence of “joint participation” between the State and the TSSAA in the TSSAA’s enforcement of its recruiting rule. The TSSAA’s board of control enforces its recruiting rule solely in accordance with the authority granted to it under the contract that each member signs.

Finally, there is no “symbiotic relationship” between the State and the TSSAA. . . . [T]he TSSAA’s “fiscal relationship with the State is not different from that of many contractors performing services for the government.” The TSSAA provides a service . . . in exchange for membership dues and gate fees, just as a vendor could contract . . . to sell refreshments at school events. Certainly the public school could sell its own refreshments, yet the existence of that option does not transform the service performed by the contractor into a state action. Also, there is no suggestion in this case that, as was the case in Burton v. Wilmington Parking Authority (1961), the State profits from the TSSAA’s decision to enforce its recruiting rule. . . .

II

. . . Under [the majority’s new] theory, . . . the combination of factors it identifies evidences “entwinement” of the State with the TSSAA, and . . . such entwinement converts private action into state action. The majority does not define “entwinement,” and

5 The majority emphasizes that public schools joining the TSSAA “give up sources of their own income to their collective association” by allowing the TSSAA “to charge for admission to their games.” However, this would be equally true whenever a State contracted with a private entity: The State presumably could provide the same service for profit, if it so chose. In Rendell-Baker, for example, the State could have created its own school for students with special needs and charged for admission. Or in Blum v. Yaretsky (1982), the State could have created its own nursing homes and charged individuals to stay there. The ability of a State to make money by performing a service it has chosen to buy from a private entity is hardly an indication that the service provider is a state actor.
the meaning of the term is not altogether clear. But whatever this new “entwinement”
theory may entail, it lacks any support in our state-action jurisprudence. Although the
majority asserts that there are three examples of entwinement analysis in our cases, there
is no case in which we have rested a finding of state action on entwinement alone.

Two of the cases on which the majority relies do not even use the word “entwinement.”

corporation that Congress created and placed under Government control for the specific
purpose of achieving a governmental objective (namely, to avert the threatened extinction
of passenger train service in the United States). Without discussing any notion of
entwinement, we simply held that, when “the Government creates a corporation by
special law, for the furtherance of governmental objectives, and retains for itself
permanent authority to appoint a majority of the directors of that corporation, the
corporation is part of the Government for purposes of the First Amendment.” Similarly,
in *Pennsylvania v. Board of Directors of City Trusts of Philadelphia* (1957), we did not
consider entwinement when we addressed the question whether an agency established by
state law was a state actor. In that case, the Pennsylvania Legislature passed a law
creating a board of directors to operate a racially segregated school for orphans. Without
mentioning “entwinement,” we held that, because the board was a state agency, its
actions were attributable to the State.

The majority’s third example, *Evans v. Newton* (1966), lends no more support to an
“entwinement” theory . . . . Although *Evans* at least uses the word “entwined,” we did not
discuss entwinement as a distinct concept, let alone one sufficient to transform a private
entity into a state actor when traditional theories of state action do not. On the contrary,
our analysis rested on the recognition that the subject of the dispute, a park, served a
“public function,” much like a fire department or a police department. A park, we noted,
is a “public facility” that “serves the community.” Even if the city severed all ties to the
park and placed its operation in private hands, the park still would be “municipal in
nature,” analogous to other public facilities that have given rise to a finding of state
action: the streets of a company town in *Marsh v. Alabama* (1946), the elective process in
*Terry v. Adams* (1953), and the transit system in *Public Util. Comm’n of D.C. v. Pollak*
(1952). Because the park served public functions, the private trustees operating the park
were considered to be state actors.6

These cases, therefore, cannot support the majority’s “entwinement” theory. Only *Evans*
speaks of entwinement at all, and it does not do so in the same broad sense as does the
majority. Moreover, these cases do not suggest that the TSSAA’s activities can be

---

6 We have used the word “entwined” in another case, *Gilmore v. Montgomery* (1974), which the majority
does not cite. In *Gilmore*, we held that a city could not grant exclusive use of public facilities to racially
segregated groups. The city, we determined, was “engaged in an elaborate subterfuge” to circumvent a
court order desegregating the city’s recreational facilities. The grant of exclusive authority was little
different from a formal agreement to run a segregated recreational program. Thus, although we quoted the
“entwined” language from *Evans*, we were not using the term in the same loose sense the majority uses it
today. And there is certainly no suggestion that the TSSAA has structured its recruiting rule specifically to
evade review of an activity that previously was deemed to be unconstitutional state action.

179

Draft — do not distribute!
considered state action, whether the label for the state-action theory is “entwinement” or anything else.

Because the majority never defines “entwinement,” the scope of its holding is unclear. If we are fortunate, the majority’s fact-specific analysis will have little bearing beyond this case. But if the majority’s new entwinement test develops in future years, it could affect many organizations that foster activities, enforce rules, and sponsor extracurricular competition among high schools—not just in athletics, but in such diverse areas as agriculture, mathematics, music, marching bands, forensics, and cheerleading. Indeed, this entwinement test may extend to other organizations that are composed of, or controlled by, public officials or public entities, such as firefighters, policemen, teachers, cities, or counties. I am not prepared to say that any private organization that permits public entities and public officials to participate acts as the State in anything or everything it does, and our state-action jurisprudence has never reached that far. The state-action doctrine was developed to reach only those actions that are truly attributable to the State, not to subject private citizens to the control of federal courts hearing § 1983 actions. . . .


The Court frequently describes state action analysis as having two prongs: first, whether “the [challenged] deprivation . . . [was] caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the State or by a person for whom the State is responsible”; and second, whether “the party charged with the deprivation . . . [is] a person who may fairly be said to be a state actor.” The nub of the inquiry is at this second step, often alternatively characterized as determining whether “there is a sufficiently close nexus between the State and the challenged action.” The Court also considers a consistent set of criteria in assessing whether this nexus exists: whether the private party is performing a public or government function; whether the government compelled or significantly encouraged the challenged action; whether the government jointly participated in the action; and whether there is symbiotic interdependence between the government and the private party. At times, however, the Court adopts a highly restrictive and formalistic stance in applying these criteria, treating them as distinct “tests” that represent the exclusive grounds for finding state action. On other occasions, the Court applies a flexible, fact-sensitive, and pragmatic approach, identifying these criteria as factors to consider in reaching an overall gestalt sense of whether the private actor should be viewed as government.

Placing the Court’s state action jurisprudence in historical perspective[, with the high point of state action corresponding to the desire to stamp out discrimination in the civil rights era,] certainly reduces some of the inconsistency in state action doctrine. . . .
Private nursing homes providing long-term care to Medicaid beneficiaries are not state actors, even though they too operate under contract with the government and make need determinations authorized by statute. Nor are private schools with which the government contracted to fulfill its statutory obligation to provide education to special needs students. Yet a private doctor treating prisoners pursuant to a contract with a prison is a state actor, as are private employers implementing government rules authorizing drug testing of transportation employees. Similarly, the United States Olympic Committee (USOC), a corporation created by federal statute and given control over U.S. participation in the Olympics as well as exclusive oversight of private amateur sports organizations participating in international competition, is not a state actor, and neither is the National Collegiate Athletic Association (NCAA). But a private organization overseeing nearly all public and private high school athletic events is, according to . . . Brentwood Academy.

These decisions demonstrate that the line of division separating state and private action remains far from straight. The continuing disagreement regarding how to conduct state action analysis is starkly apparent in a comparison of the Court’s two most recent state action decisions, Sullivan and Brentwood Academy. . . . Obviously, the two cases differ in their underlying facts and particularly in the relationship of the private entities involved with the government. While the government created and closely regulated the workers’ compensation system, the duty to provide benefits lay by statute with employers; the government did not participate in the private insurers’ refusals to pay benefits other than to authorize such refusal prior to a hearing. By contrast, the Court emphasized the numerous connections between the Tennessee Secondary School Athletic Association and the state in finding the Association to be a state actor, in particular the fact that a substantial majority of Association members were Tennessee public schools and that public school officials “overwhelmingly perform all but the purely ministerial acts by which the Association exists and functions.”

Even so, a striking difference in analytic style exists between the two decisions. Sullivan is in many ways the apogee of a formalistic state action inquiry, with the Court insisting upon state involvement in “the specific conduct of which the plaintiff complains” and not just background connections to the private entity. The Court also reiterated other limiting rules and maxims developed in earlier cases that significantly narrow the scope of the state action inquiry, such as that “[t]he mere fact that a business is subject to state regulation” or “the mere approval or acquiescence of the State” in a private entity’s actions do not create state action. In Brentwood Academy, by contrast, the majority opinion took an avowedly flexible, pragmatic, and situation-specific approach, focusing on the “practical certainty . . . that public officials will control operation of the Association”; the Court stated that “[w]hat is fairly attributable [to the state] is a matter of normative judgment, and the criteria lack rigid simplicity.” Moreover, it was precisely the background connections between the Association and the state—in the Court’s words, “pervasive entwinement”—rather than state involvement in specific Association decisions that formed the basis for the Court’s finding of state action.

Quite clearly, such differing approaches make hazardous any confident prediction of the effect of ever-increasing privatization on state action doctrine. Yet notwithstanding the
seeming trend toward a more flexible and perhaps expansive state action inquiry in *Brentwood Academy*, the pervasive thrust of the Court’s recent decisions strongly suggests that privatization is likely to result in a denial of state action. Nor is this outcome surprising when the rules of current state action doctrine are compared to the practical realities of privatization.

One key factor leading to denials of state action is the Court’s restriction of what satisfies the public function test to functions that are traditionally and exclusively performed by government. Taken literally, this formulation excludes most of the tasks commonly performed by government today. Private organizations played a central role in distributing assistance to the poor prior to the New Deal, while private prisons and private police similarly have a longstanding pedigree. More importantly, the Court has relied on this narrow formulation to reject the claim that private entities are performing a public function for state action purposes when they are fulfilling duties ultimately borne by the government. Instead, private provision of services on the government’s behalf only constitutes a public function if the services are ones that the government is required to provide directly.

The overall effect of this approach is to remove one of the strongest arguments for finding state action in contexts of government privatization. Admittedly, instances exist where the government itself must remain directly responsible for the duties involved. *West* appears to be such a case; there, the private doctor held to be a state actor was fulfilling the state’s Eighth Amendment obligation to provide adequate medical care to prisoners, an obligation that by virtue of the public-private divide runs only against the government. But governments rarely are subject to such direct provision requirements. Instead, it is expected that governments will often meet their obligations to provide services by relying on private providers, and the activities involved are frequently far removed from historically core government functions.

As a consequence, the only grounds generally available for finding state action will be the government’s connections or interaction with the private entity. Here, the obstacle to finding state action in government privatization contexts is the requirement, reiterated in *Sullivan*, that government participation must be present in the specific action being challenged; absent such participation, only government compulsion or formal control of the private actor will suffice. Public control of a private entity or participation in specific actions is unlikely to be present in a great many instances of privatization. As decisions such as *West* demonstrate, situations will arise where a private actor operates within and is an integral part of an institution controlled by the government. But more often privatization means handing over program implementation in its entirety to separate private entities or, increasingly, giving private actors responsibility for running what formerly were public institutions, such as a prisons, schools, or hospitals. This is particularly true in social welfare programs, where the government relies on private entities to provide services to large numbers of beneficiaries. Private providers regularly make day-to-day decisions on their own; public involvement is limited to issuing general requirements that providers must meet (through regulations or contractual terms) and conducting periodic reviews. Indeed, taking advantage of private freedom from public control—identified as exemption from civil service provisions, bureaucratic rules,
substantive programmatic requirements—is often the main impetus behind efforts to privatize.

The question then becomes to what extent 

_Brentwood Academy_ has altered the Court’s prime insistence on government involvement in specific challenged acts. Clearly, background involvement that reaches the level of “pervasive entwinement” is now a potential ground for finding state action. However, in holding that such pervasive entwinement was present, the Court placed particular weight on the participation of public officials in the Association, noting that these officials “do not merely control but overwhelmingly perform all but the purely ministerial acts by which the Association exists and functions . . . .” Again, instances where government officials control or constitute private entities are hardly the norm; moreover, _Brentwood Academy_ went out of its way to indicate that less extreme forms of government involvement might not suffice, stating that government purchases of contract services “do not convert the service providers into public actors.” Notably, in making this point _Brentwood Academy_ cited _Rendell-Baker_ with approval. In _Rendell-Baker_, the Court held that a private school was not a state actor, notwithstanding that the school received nearly all of its students through contracts with local or state authorities, was heavily regulated by the state, and relied on state payments for the vast majority of its operating budget. If such extensive government involvement does not suffice to create state action, claims of pervasive entanglement will rarely succeed.

What frequently is present in instances of privatization is a formal delegation of authority from the state to the private entity, whether by contract or by governing statutes and regulations. However, these delegations have little significance under current state action doctrine. _Brentwood Academy_ is again a good example. There, the majority noted that for twenty-five years Tennessee maintained a regulation expressly designating the Association as the entity responsible for supervising all high school athletics in the state. Interestingly, however, the Court treated this designation as further evidence of the state’s entwinement and close relationship with the Association, instead of suggesting that delegation of authority may on its own form a basis for finding state action. To the extent that consideration of whether the government has delegated authority exists, it usually arises in determining whether the public function test is satisfied. But in this context, the analysis focuses on the nature of the task being delegated rather than on the act of delegation, with little consideration given to how the delegation may affect the private delegate’s powers. Indeed, if anything, the Court’s decisions indicate that a broad delegation of power actually serves as a basis for denying the existence of state action, because a private delegate’s exercise of independent judgment and discretion is taken as strong evidence that significant government control and involvement is lacking.

Hence, it seems quite unlikely that state action will be found in most instances of government privatization, at least under current state action doctrine. . . . No doubt, the Supreme Court will clamp down when it perceives an effort by government to evade its constitutional obligations. But privatization decisions rarely exhibit clear evidence of constitutional bad faith. Is privatizing to avoid the cumbersome, bureaucratic processes that come with government an illegitimate effort to bypass procedural due process, or is it
instead an admirable attempt to inject innovation and beneficial flexibility into government programs?

Additional support for the conclusion that state action will rarely be found comes from the transformation over the last two decades in the animating concerns of state action doctrine. *Brentwood Academy* aside, few of the Court’s state action decisions even identify—let alone emphasize—the importance of ensuring that exercises of government power do not escape constitutional constraints as an underlying imperative of state action doctrine. Far more common of late is for cases to underscore the values at stake in *not* finding state action, expressing concern that state action holdings threaten individual autonomy, federalism, and the regulatory prerogatives of elected government. Particularly given the current Court’s emphasis on protecting these values in other jurisprudential areas, it seems unlikely that it will adopt a more expansive approach to state action in response to increased privatization.
§ 3. State Action Applied: Private Police


INTRODUCTION

For most lawyers and scholars, private security is terra incognita—wild, unmapped, and largely unexplored. Criminal procedure, the branch of constitutional law that aims to regulate the conduct of the police, is a well-developed field of jurisprudence and of legal scholarship. The complaint is heard increasingly often that contemporary criminal procedure is too well developed—too ornate, too complicated, too far removed from first principles. No one makes that complaint about private security law. Indeed, no one even speaks of “private security law,” for the very phrase suggests a unified and specialized body of rules that does not exist. Instead, private security personnel find their conduct governed by a hodgepodge of private contract provisions, state and local regulations, and tort and criminal law doctrines of assault, trespass, and false imprisonment. On those rare occasions when private security employees become “state actors,” constitutional criminal procedure gets added to the mix. Partly as a result of this disarray, legal scholars have tended to ignore private security.

The neglect is increasingly indefensible. The private security industry already employs significantly more guards, patrol personnel, and detectives than the federal, state, and local governments combined, and the disparity is growing. Increasingly, private security firms patrol not only industrial facilities and commercial establishments but also office buildings, shopping districts, and residential neighborhoods. Private policing poses no risk of supplanting public law enforcement entirely, at least not in our lifetime, and it is far from clear to what extent the growing numbers of private security employees are actually performing functions previously carried out by public officers. Still, if criminal procedure scholars continue to focus exclusively on the public side of law enforcement, our work is likely to become of steadily more marginal importance.

More importantly, ignoring private policing has impoverished our thinking about the public police and about constitutional criminal procedure. Private security firms offer tangible evidence about what some people want but are not receiving from public law enforcement, and what the public may increasingly pressure police departments to provide. The legal regime governing private security, moreover, is strikingly similar to the legal regime that many reformers have advocated for public law enforcement: deconstitutionalized, defederalized, tort based, and heavily reliant both on legislatures and on juries. It therefore offers tantalizing opportunities to test some of the most persistent objections to modern criminal procedure law. And because maintaining order and controlling crime are paradigmatic governmental functions, private policing offers a unique and underused vantage point for reexamining the public-private distinction and the state action doctrine. Finally, police privatization furnishes an opportunity to
reconsider the focus of constitutional law on negative obligations of government, and the
overwhelming focus of constitutional criminal procedure on fairness to individual
criminal defendants. The dramatic spread of policing-for-hire may require rethinking, for
example, what it means to guarantee all citizens, regardless of wealth, the equal
protection of the laws.

[After, in Part I, describing the private police and the legal rules governing it, and, in Part
II, describing the history of private policing,] I turn in Part III to perhaps the most
obvious challenges that private policing poses for criminal procedure jurisprudence and
scholarship: the unresolved and possibly irresolvable dilemmas it creates for the state
action doctrine in criminal procedure. I consider here not only the problems that private
policing presents for the concept of state action, but also the lessons it may offer. Because
it touches so strongly our ambivalence about the public-private divide, the private
security industry places great strain on the doctrine that constitutional rights constrain
only action attributable to the government. The manner in which courts have responded
to that strain throws light on the centrality of symbolism in our understanding of the
police and on the inevitability of a certain kind of formalism in state action law. Along
with the formalism, I suggest, comes an unavoidable amount of arbitrariness and
incongruity. Once recognized, however, those defects can perhaps become strengths.

Ultimately, I argue that police privatization raises challenges more important than
determining whether to view security guards as state actors. Part IV addresses two of
these challenges. The first is to learn more about the dimensions of the private police and
about how they operate. This information is necessary for a sensible assessment of the
rules governing private security personnel and could help test any number of proposals
for reforming public law enforcement. The second challenge, more difficult than the first,
is to reconsider whether the state owes all its citizens some minimally adequate level of
police protection. An affirmative entitlement to policing may seem far from the concerns
most immediately raised by the vaguely threatening private guards that increasingly
patrol our daily lives. The chief task of this Article, however, is to demonstrate that
private policing raises questions far deeper, and far more interesting, than we generally
have recognized.

The private security industry forms part of at least two larger phenomena in
contemporary criminal justice. The first is criminal justice privatization, a trend that
includes private policing as well as private prisons and private adjudication. The second
is the “pluralizing of policing”—that is, the partial displacement of public policing not
only by private security personnel, but also by community volunteers. Although much of
what follows will bear on those larger trends, I will have little to say about them directly.
This is not because they are uninteresting or unimportant, but rather because the private
police, for reasons I hope to make clear, pose special challenges and opportunities for
judges and scholars.
I. A PRIVATE POLICE PRIMER

B. Private Security Law

... Criminal procedure law—the vast set of interrelated constitutional doctrines that regulate the day-to-day operations of police officers throughout the United States—has almost nothing to say about the activities of private security guards. That law consists chiefly of the Fourth, Fifth, and Sixth Amendments, made applicable to the states through the Due Process Clause of the Fourteenth Amendment, along with the Supreme Court’s elaborate efforts to implement those provisions through rules of evidentiary exclusion and the restrictions on interrogations imposed by *Miranda v. Arizona* and its progeny. All of this has been deemed applicable only to government action, not to private conduct. As a result, the “criminal procedure revolution” of the past half century has left private security largely untouched. Private searches fall outside the coverage of the Fourth Amendment, and evidence they uncover is almost always admissible; similarly, suspects interrogated by security guards are not entitled to *Miranda* warnings and generally do not receive them.

The main legal limitations on the private police today are tort and criminal doctrines of assault, trespass, and false imprisonment—variants of the same doctrines that once defined the principal boundaries of permissible public policing. Unless the owner has given consent, a security guard’s search of private property will generally constitute a trespass. And arrests or detentions not authorized by state law generally will expose a security guard to civil and criminal liability for false imprisonment and, if force is involved, for assault.

Private security companies eager to appear unthreatening often stress that their personnel are limited to the search and arrest powers of ordinary citizens. It is a mistake, though, to make too much of this limitation. In the first place, it is not always true that security guards have only the powers of ordinary citizens. Many private guards, for example, are “deputized” or otherwise given full or partial police powers by state or local enactment, and most states have codified a “merchant’s privilege” that allows store investigators, and in some instances other categories of private security personnel, to conduct brief investigatory detentions that would be tortious or criminal if carried out by ordinary citizens.

In the second place, the arrest powers of ordinary citizens in most states are not strikingly different, in some significant respects, from those of police officers. Officers can execute arrest warrants; private persons generally cannot. But the vast majority of arrests are made without a warrant, and the arrest powers of officers and civilians in that circumstance are relatively narrow. An officer, as a general matter, may arrest anyone he or she has probable cause to believe has committed a felony, and anyone who commits a misdemeanor in the officer’s presence. A private citizen typically may also arrest for a misdemeanor committed in his or her presence, and for a felony he or she has probable cause to believe the arrestee has committed—as long as the felony has in fact been committed, by the arrestee or by someone else.
The difference between the arrest powers of police officers and private citizens, then, comes down to this: a private citizen, unlike a police officer, is liable for false arrest if he or she arrests someone based on probable cause to believe the arrestee committed a felony, but it turns out that the felony had not actually been committed. Neither is liable if it turns out the felony has been committed, but not by the defendant.

This distinction between the arrest powers of officers and those of private persons was not always drawn, and not all states draw it today. Even when the distinction is drawn, moreover, there is less to it than meets the eye. For if the defendant’s good faith does not defeat tort liability for false arrest, it generally does preclude punitive damages and can mitigate actual damages, not to mention dampen the case’s jury appeal. As a result, the recovery in a false arrest case brought successfully against a private person who acted reasonably and with probable cause is likely to be quite low—which may explain why such cases appear to be rare. Successful criminal prosecutions in such instances appear virtually nonexistent.

When it comes to arrests, the real differences between police officers and private persons lie not in their powers, but in the consequences that ensue when those powers are exceeded. On the one hand, police officers who act reasonably and in good faith generally are immunized from tort or criminal liability for false arrest—even if they make a mistake about what is “objectively legally reasonable.” Private persons, including security guards, typically lack this protection. On the other hand, evidence generated by an illegal arrest by a police officer is, as a general matter, inadmissible against a criminal defendant; the fruits of private illegality are not similarly excluded. Instead of two separate doctrines of arrest, we have two separate remedial systems: tort actions for violations committed by private police, and suppression motions for evidence obtained illegally by public law enforcement.

Actually, even the remedial systems overlap, because tort actions can be brought against public officers as well as private security personnel. True, public officers can claim qualified immunity when they operate in good faith, but it is unclear how much the lack of such immunity hurts private security firms, given the practical obstacles to recovery against a defendant who has acted in good faith. Moreover, in at least one important respect tort suits are easier to bring against public police officers than against private security personnel: suits alleging constitutional violations by officers acting “under color of state law” can proceed under 42 U.S.C. § 1983, which entitles victorious plaintiffs to recover not only damages but attorneys’ fees as well.

None of this is to suggest that there are no significant legal distinctions between the powers of public and private police. The public police obviously have some well-defined powers that private security personnel lack. This is particularly true with regard to searches. Law enforcement officers but not private citizens can apply for and execute search warrants and electronic surveillance orders, and in many circumstances the Supreme Court has granted police officers, but not private citizens, broad powers to search without a warrant. In addition, police officers but not private citizens generally are empowered to command the assistance of bystanders. There are also differences in the
consequences that attach to a failure to submit to arrest. Resisting even a lawful citizen’s arrest typically is not a crime, although it frequently will be tortious. In contrast, most states now criminalize resisting an arrest by a law enforcement officer even when the arrest is illegal.

But the greatest differences in the powers of the public and private police are implicit. As we will see, they flow from the social understandings that surround the role of the public law enforcement officer. Before tracing the roots of those understandings, however, we need to survey their current manifestations.

C. Policing and the Public-Private Distinction

Thinking about private policing unavoidably involves thinking about the public-private distinction more generally. This is partly because policing—peacekeeping, property protection, and law enforcement—touches on deep and contradictory intuitions regarding the proper allocations of responsibilities between the public and private spheres. On the one hand, peacekeeping, property protection, and law enforcement are often considered the clearest examples of functions that are essentially and necessarily public, and therefore essentially and necessarily the job of government. The idea here—loosely shared by John Locke and Max Weber, and latterly by Robert Nozick and Ronald Reagan—is that the very point of government is to monopolize the coercive use of force, in order to ensure public peace, personal security, and the use and enjoyment of property. (Hence the classic description of the libertarian ideal: “the night watchman state.”) One reflection of this idea is the common notion that it is wrong to “take the law into your own hands.” Another is the view, taken as self-evident by the Supreme Court, that “the most basic function of any government is to provide for the security of the individual and of his property.”

On the other hand, private policing can easily be understood as the natural product of three paradigmatically private functions. The first is self-defense, widely viewed as an inherent right, particularly in America, just as “taking the law into your own hands” is seen as obviously wrong. The second is economic exchange, the “free market” that, as we will see, transformed the eighteenth-century constabulary from a civic duty to a specialized form of employment. The third is the use and enjoyment of property, generally thought to include the right of owners to place conditions on those invited onto their property.

Not surprisingly, therefore, the recent growth of private policing has elicited conflicting reactions, reflecting broader conflicts about government in general. The trend has been applauded on three different grounds. First, private policing has been welcomed as more flexible than traditional, public law enforcement. Private guard companies, unlike public police forces, are free from civil service rules, reporting requirements, and the range of other rules characteristically imposed on government agencies. In addition, private companies lack many of the bureaucratic traditions that may handicap the effectiveness of the public police, and they are often free from the constraints associated with a unionized workforce. These factors allow guard companies to act in ways in which government either cannot or will not.
Second, private policing has been celebrated as more *accountable* than its public counterpart. Unlike public police forces, private guard companies have to answer to the discipline of the market. In the words of one enthusiast, a privately employed police officer inevitably “recognizes the importance of establishing positive relationships with the consumers of the service and develops innovative approaches to community problems,” whereas “the public police are paid through the compilation of public taxes and are, therefore, answerable to every business and citizen in the city but are not accountable to them.” Private policing thus can be understood in part as a reaction against the excessive independence and insularity of modern police forces—an answer to the common complaint “that a police force should be responsive to those policed, while autonomous professionals . . . have a notorious tendency to believe that they know what is ‘really’ best for the clients, and therefore what the client ‘ought’ to want.”

Third and finally, private policing has been thought beneficial because it *empowers* those it protects. Unlike the first two advantages of private policing, this third one does not have to do with the improved performance of those doing the policing, but rather from the effect on those who hire them. The argument here has two strands. The first is *individualistic*: it appeals to the notion that every citizen should take responsibility for his or her own protection, that it is ultimately enfeebling to depend on the government for protection. The second is *communitarian*: the idea here is that arranging for private policing generally entails a more or less voluntary association of residents or business owners that, in the process of providing joint security, also builds social capital—which itself can help reduce crime. The first strand of the empowerment argument thus envisions private security as the natural outgrowth of individual self-defense; the second strand sees it as a commercial variant on citizen patrols.

Each of these claimed advantages of private policing, of course, has a flip side. For example, the greater flexibility of private guard companies stems from their freedom from regulations and traditions that have developed largely because they were thought necessary to control the uniformed, armed, quasi-military forces patrolling our streets. Where some see the greater flexibility of private policing, others see the threat of policing that is *uncontrolled*. Indeed, the most persistent complaint about private guard companies—a complaint that, we will see, has deep roots in the nineteenth century, and that today serves as a perennial focus of television exposés—is that they are insufficiently regulated.

The frequency with which this complaint is raised serves as a reminder that the supposed accountability of private policing has a troubling side as well. Private companies are thought more accountable than government because they answer to their particular customers instead of to the general public. But this is not clearly to everyone’s advantage. In particular, those who come into contact with private guards but do not help to pay for them may not welcome the fact that such guards are accountable exclusively to their customers. And even some of the customers, when they venture outside their own territory, may wish that the various uniformed patrol personnel they encounter were less *proprietary*, more answerable to the general public.
More broadly, there are grounds for doubting that market forces will deliver optimum levels of police protection. In the lingo of economics, policing gives rise to three sorts of externalities. First, there is the free-rider problem: if my neighbors pay for a private patrol car to drive by their homes periodically, burglars may be scared away from my house, too. Because there is no practical way to limit all the benefits of patrol service to those who pay for it, markets might be expected to supply a suboptimal level of patrolling. Second, there is the problem of displacement. The patrol car in my neighborhood may push burglars into surrounding neighborhoods, increasing crime there while decreasing it where I live. Indeed, some industry executives boast that private patrols, once established in an area, create further marketing opportunities by moving crime to adjacent, underpatrolled areas. Alarm systems, advertised by lawn placards and window decals, may have a similar effect within neighborhoods. Third, there is the problem mentioned above: policing protects some people by interfering with other people, and there is no obvious way to require those who are protected to pay for the burdens imposed on those they are protected from. All these externalities can be anticipated to warp private expenditures for police protection away from what economists would consider socially efficient. One might expect the private sector to overfund crime control strategies with large negative externalities, and to underfund strategies with large positive externalities: more money for home alarms than for private patrols, and more for private patrols than for antipoverty foundations.

Nor, finally, are the broad effects of private policing on its employers unambiguously positive. Public law enforcement has been celebrated as “socializing” the coercive use of force. To the extent that private policing is seen as an extension of self-defense, it can be—and has been—deplored as a retreat from that process. On the other hand, to the extent that private policing is seen as a commercial variant on citizen patrols, the commercial aspect may be thought to eliminate much of the civic value. And not everyone is enthusiastic about citizen patrols even in their pure form; American history gives ample cause for disquiet about amateur law enforcement. Indeed, as we shall see, the history of policing, both in England and in America, throws doubt on most easy answers to the problems presented by private policing.

II. A SHORT HISTORY OF PUBLIC AND PRIVATE POLICING

. . . [The history of policing shows that p]rivate organizations have a long tradition of filling perceived gaps in the policing services provided by government. I argue that there is reason to believe this is happening again today, and that the demand for private security should not be discounted as irrational.

Other lessons [from history] pertain to the nature of policing. Many people today view crime control and order maintenance as inherently public functions—perhaps the paradigms of jobs assigned traditionally and uncontroversially to government. Others regard the preference for public policing as an arbitrary, possibly transient fancy of the postwar decades. History throws doubt on both views. Policing as a concept has been remarkably malleable, and no part of the job has ever been monopolized by government. But the malleability of policing can easily be exaggerated; the continuities in organized policing since the early nineteenth century have in some ways been as striking as the
changes. Moreover, if the police functions deemed appropriately public have varied over time, the sense that some aspects of order maintenance and crime control should be kept in public hands turns out to be both old and durable. In the context of policing, at least, the public-private distinction is both more flexible and more confining than sometimes thought.

Finally, history offers humbling lessons about the limited salience of legal rules in constructing the police as a social institution. The special powers of law enforcement officers have never been well defined, but they have always been assumed to exist; the police have been defined more by culture than by law. I argue that understanding this fact is the first step to thinking more sensibly about the state action problem in criminal procedure, and possibly about the state action doctrine more generally. The special role of the state, like the special role of police officers, may be part of our cultural baggage, and harder to jettison than often imagined.

[From the Middle Ages through the 19th century, law enforcement was a public-private affair. “Constables and watchmen were not paid for their service, and ‘police duties were the duties of every man.’” 17th-century treatises “drew no sharp distinction between the powers of constables and those of . . . ‘private persons’: both could arrest individuals ‘probably suspected of felonies, but only if a felony had in fact been committed.’” This system of mandatory community service, in which those with money hired deputies to take their place, worked poorly. But constables and watchmen were not relied on to arrest criminals: victims were responsible “for catching the offenders, turning them over to the authorities, and then prosecuting the cases.” The state reinforced this private system by offering rewards—a system that bred a class of professional law enforcer called the “thief-taker,” who was often effective in fighting crime but also became notorious for colluding with the thieves themselves.

[In the 19th century, professional police forces grew in popularity, both in England and in America. They still often got their funding from private sources. From an initial exclusive focus on prevention, they gradually increased their investigative functions. “By the 1860s, the so-called ‘old police’ of unpaid constables and watchmen had been replaced throughout the country by ‘new police’—professional, full-time employees.”

[American law enforcement diverged from English models from the mid-19th century onwards. “[P]rivate law enforcement took on dramatic new forms in America.” First, particularly in the South and West, there was vigilantism. Second, in the East and Midwest, there was “policing for hire.” These private policemen were often hard to distinguish from their public counterparts, but by midcentury the public-private distinction became sharper: “one begins to find judicial opinions refusing to enforce contracts providing for private compensation to public law enforcement officers, on the ground that a police officer ‘is not the agent or employee of the private prosecutor, but the minister of the law, doing the work of the public.’” Still, “public policing remained an urban phenomenon.” Where public police protection was absent, there emerged the “‘company police’: forces of guards and detectives hired and supervised by railroads and industrialists to protect their own property and empowered as police officers by the
There also emerged “the national private police agency, epitomized in the late nineteenth century by the Pinkerton agency.

Allan Pinkerton started out offering “uniformed night watch service to Chicago businesses.” After the Civil War, his agency’s main source of income was spying on the employees of railroads and express companies to test their honesty. Starting in the 1870s, the agency’s work shifted to “protecting clients against growing labor unrest . . . . Of course this was not all it did; Pinkerton had become by this time America’s de facto national law enforcement agency, albeit an agency whose services were available only to those of ample means. . . . Hostility to private policing mounted during the second half of the nineteenth century, fueled by periodic stories of malfeasance and by a growing notion that the responsibility for peacekeeping should not be placed in private hands.” Congress investigated the Pinkerton agency and its competitors in the 1890s, after a disastrous labor disturbance in Homestead, Penn. But private detective firms continued to grow in the early 20th century. The Pinkerton agency itself focused more on crime fighting than on strike services (it abandoned labor relations services altogether after public scrutiny in the New Deal era), and gradually, “the investigative and managerial techniques of private detective firms were adopted by the public sector,” including the FBI. By the mid-20th century, “private policing was considered an anachronistic institution that had withered away in response to the growth of the “new police.””

Some of what looked like retreat, however, was actually redeployment: Pinkerton and its rivals were turning from detective firms that also provided guards into security guard companies that offered detective services on the side. Inverting the transformation undergone by public police agencies, which had moved from an emphasis on preventive patrol to an emphasis on detection and apprehension, the private sector moved from the detection and apprehension business to the preventive patrol business.” In the postwar years, scholars started to examine policing; they concluded that public policing was the wave of the future and that private policing would inevitably shrink. But by the early 1980s, it was clear that the private security industry was rapidly expanding. “[P]rivate security guards now greatly outnumber uniformed law enforcement personnel and probably will outnumber them even more dramatically in coming years.”

F. Provisional Lessons

1. The Roots of Police Privatization

At least three different explanations have been offered for the stunning growth of private policing in recent decades. The first, and perhaps the simplest, attributes it to ideological shift: police privatization, in this view, is part of a broader shift of resources and responsibilities away from government and toward the private sector. Police privatization in the United States, however, predates the broader national trend toward privatization by at least a decade.

A second, widely circulated explanation for the recent growth of private police was first offered fifteen years ago by Canadian criminologists Clifford Shearing and Philip Stenning. They attributed that growth to a significant increase in the amount of “public”
activity taking place on what they called “mass private property”: large, privately owned facilities such as shopping malls, office buildings, housing complexes, manufacturing plants, recreational facilities, and university campuses. “The modern development of mass private property,” they argued, “has meant that more and more public life now takes place on property which is privately owned.” This in turn, they contended, has led to a steadily increasing role for private security in the provision of public order, because the public police, whose patrol responsibilities traditionally have been limited to public property, generally have lacked both the resources and the desire to maintain order on private property, and because private owners, in any event, have preferred to keep control over the policing of their property.

On its face this theory does a better job than ideological shift in explaining the timing of the recent growth in private policing, but it has two significant limitations. The first is a lack of convincing empirical support. Although Shearing and Stenning suggested that evidence of the growth of mass private property “surrounds every city dweller,” they acknowledged there was little data to support their claim. More importantly, much of the most visible growth of private security in recent years has occurred on public property, funded by groups of businesses or homeowners dissatisfied with public police protection. That dissatisfaction, in fact, appears itself to have contributed to the growth of at least some highly visible forms of mass private property, including gated residential communities and privately controlled shopping and recreational complexes.

This leads us to the most widespread explanation for the dramatic postwar growth in private security: the failure of public law enforcement to provide the amounts and the kinds of policing that many people want. This explanation has some intuitive plausibility, and it also comports with past experience. As we have seen, for over two centuries privately paid entrepreneurs in both Britain and America have been filling gaps in the police protection offered by public law enforcement. Private police today, moreover, tend at least in broad outline to do the kinds of things that public police departments are faulted for not doing: patrol visibly and intensively, consult frequently with the people they are charged with protecting, and—most basically—view themselves as service providers.

Ironically, the labor-intensive patrolling provided by private security firms may depend on the same factor that causes many of the industry’s problems: low wages. Personnel expenditures, more than anything else, limit the ability of police departments to increase their “visible presence,” and security guards cost far less than police officers. Studies of the private security industry regularly blame the meager wages paid to security guards for “marginal personnel” and high turnover, and blame high turnover for the minimal training guards receive. But a cheap pair of eyes may be just what many customers want. Private police may be the modern-day equivalents of eighteenth-century watchmen: poorly paid and marginally competent, but competent enough to perform a service.

Plainly not all of the growth in private policing has been driven by preexisting demand. Some of the growth has been fueled by crime-related anxieties that the marketing arms of private security firms, along with escalating coverage of crime by the news media, have
helped to amplify and to sustain. In addition, the privatization of policing has its own momentum; the process tends to sustain itself, as people get used to the idea of relying on private security. One way this happens is through the increased willingness of judges and juries to award tort damages for a landowner’s failure to provide adequate private policing; another is through the decreased incentive that clients of private security firms have to support large public expenditures for police protection; a third is through the new demand that can be created when private patrols in one area move crime to adjacent, underpatrolled areas.

Still, it seems a mistake to write off the demand for private policing as irrational. Private policing would not have ballooned the way it has in recent decades if customers were unwilling to pay for it, and today’s private guard companies—much like the thief-catchers of Georgian England or the national detective agencies of late-nineteenth-century America—provide an indication of the kinds of police services for which the government has left demand unmet.

Obviously, not all demands should be met, particularly not by government. The detective agencies of the Pinkerton era provide a good example: they certainly demonstrated that corporate America wanted more help breaking strikes than the government was providing, but that did not necessarily mean that the government should start providing that help. Today, some part of the demand for private security services is a demand for keeping certain kinds of people—typically poor or members of racial minorities—out of the business districts, amusement parks, and residential areas that private guards are hired to patrol. Not only is this a demand the government has no business helping to meet; it is one that most people today believe that government should help suppress.

But private security firms also serve more benign interests, interests that a wide spectrum of Americans believe the government itself should pursue more aggressively. Chief among these are protecting people against serious crime—and protecting them against the fear of serious crime, in part by preserving public order. These are of course among the principal goals of current efforts at police reform. Among the simplest lessons that police privatization offers to students of public law enforcement are not to dismiss these efforts, and not to underestimate the breadth and the depth of the dissatisfaction to which they respond.

2. The Nature of Policing

The history of policing poses serious challenges for the view that policing is an inherently public function. It has never been an entirely public function, and only in the second half of the nineteenth century did a relatively clear dividing line emerge between public and private policing. The world of the 1990s, of course, is different from the worlds of the 1790s and 1890s; perhaps a case can be made that today’s technological and social realities make a public monopoly on policing particularly important. But those who describe law enforcement as an essentially public function rarely attempt to make such a case; instead, they typically suggest that private policing is inherently inconsistent with the liberal state. The history of policing makes that view difficult to defend.
Past experience also poses challenges for other essentialist understandings of policing—for example, that policing is first and foremost about preventive patrol, or first and foremost about catching criminals, or that one of these functions is traditionally public and the other traditionally private. In the nineteenth century, police departments specialized in patrol work, often leaving detective work to the market. In the mid-twentieth century, police departments changed their principal claim of expertise—if not their principal area of practice—from prevention to apprehension, and private security firms, in reaction, made the opposite transition. Congressional reformers in the 1890s expressed their chief dismay at the use of private police for maintaining order; by the 1930s, property protection struck the La Follette committee as the one proper function of private security.

Nevertheless, the malleability of public expectations about policing should not be overstated. The notion that protection from crime should be publicly provided is centuries old and has been a constant force in the development of modern police forces, even if the precise implications of that notion have rarely been clear. It thus would be a mistake to dismiss current concerns about private policing as simply reflecting an ignorance of the past.

One can also err by making too much of the distinction between patrol and detective work. This is not to say that the distinction is unimportant or unhelpful—far from it. It is hard to understand the changing roles of public and private police in the nineteenth and twentieth centuries without differentiating between preventive patrol work on the one hand and detection and apprehension of criminals on the other. And both the public and private sectors of policing sharply segregate patrol and investigative personnel, and have done so for well over a century. Still, the distinction can be deceptive because patrol and detection are, to a great extent, unavoidably intertwined. Rarely has any important police organization, whether public or private, maintained an exclusive focus on patrol or investigation over any significant length of time. The Bow Street Runners branched out into patrol work, London’s Metropolitan Police and its American copies all formed detective divisions, Pinkerton’s nineteenth-century detective agency earned much of its money providing guards, and today “many, and probably most, guard companies provide some form of investigative service.”

In practice there is substantial overlap between patrol and detective work. Good patrol officers, whether public or private, necessarily engage in a large amount of detection: they must be alert to signs of criminal activity and adroit at determining the nature, causes, and possible solutions to the many low-level disturbances they encounter. In addition, because they are the first on the scene, patrol officers are often in the best position to do detective work of the more traditional, “crime-solving” sort—and, in fact, much of the most important crime solving is carried out by officers patrolling their beats. Similarly, much of the work performed by detectives amounts to covert patrol work. This was true of the Pinkerton agency’s industrial spies, and it is true today, for example, of vice detectives and narcotics investigators. Detectives need patrol skills, just as patrol officers need detective skills.
And police organizations benefit from providing both services. Not only are there economies of scale, but each operation helps to give the other credibility. Patrol officers are more likely to be feared, respected, and hired if they are part of an organization with a reputation for tracking down criminals. John Fielding and Allan Pinkerton both relied on this principle in building up their patrol business. Both also relied on a converse effect: patrol officers, highly visible by design, function as walking advertisements for investigative services.

3. Legal Roles and Legal Rules

For legal scholars, the most humbling lesson offered by the history of public and private policing is that the parameters of the police officer's role have not been set exclusively by the legal rules, and perhaps not even principally. One sign of this is the relatively low amount of attention paid throughout the nineteenth century, when modern police forces emerged and matured, to the precise powers of police officers. This topic never received anything approaching the care given to questions about what the police should wear and how they should be supervised. The authority of the police has been defined less by law than by culture and by politics. Even today, as we have seen, the arrest powers of police officers are, as a practical matter, not strikingly different from those of citizens and private guards.

Of course the police have important powers other than the power to arrest. Some of those other powers—for example, the ability to detain a suspect without carrying out a formal arrest—have been explicitly granted to some private security personnel, but other private security personnel have no special powers beyond those of ordinary citizens. But even when dealing with powers other than arrest, the legal significance of the divide between public and private police has always been hazy. The vagueness of police prerogatives is often seen as a modern phenomenon, but in fact the general powers of law enforcement officers have never been very clear. Blackstone, for example, noted that constables and other peace officers were “armed with very large powers, of arresting, and imprisoning, of breaking open houses, and the like,” but he left many of these powers uncataloged—suggesting it was “perhaps very well” that the officers generally did not know the extent of their own powers, “considering what manner of men are for the most part put into these offices[.]” And few if any grants of police powers to private constabularies in the eighteenth and nineteenth centuries specified the precise powers being granted.

Nonetheless, the idea that the public police have unique and well-defined powers has long played an important role in the public perception, and the self-perception, of private police forces. When privately employed guards have been “deputized” or “commissioned” with the “full” powers of public police officers, this has helped to legitimize their authority—despite persistent ambiguity about what those powers entail. Conversely, private security firms without such commissions have legitimized their operations, and the relative absence of public oversight over those operations, by appealing to their lack of “police powers.” These firms in some ways have the best of both worlds: the uniforms their officers wear allow them to borrow the symbolic authority of the public police, while their formal status as private citizens exempts them
from the legal requirements and extralegal expectations imposed on public law enforcement officers.

This is not just a case of industry obfuscation coupled with popular naiveté. As we shall see, the idea that police officers exist as a separate, clearly identifiable category, with unique, clearly identifiable powers, has exercised a powerful hold on courts and legislators as well as the public. “We live by symbols,” and “few symbols are better recognized than the lawman’s badge.”

III. PRIVATE POLICE AND THE PROBLEM OF STATE ACTION

The persistent strength of law enforcement symbolism can be seen nowhere more vividly than in judicial responses to the difficulties private policing creates for the state action doctrine—the general principle that constitutional rights, at both the federal and state levels, operate only against the government. In criminal cases, the state action doctrine has rendered the Fourth Amendment exclusionary rule, the *Miranda* protections, and the underlying guarantees of the Fourth, Fifth, and Sixth Amendments, inapplicable to investigative activity carried out by private citizens. Private policing poses challenges for the state action doctrine because it straddles the divide between ordinary private citizens—a concerned neighbor or vigilant storekeeper—and uniformed police officers. Are private security personnel truly private in the sense pertinent to state action?

As we will see, a principled answer proves elusive. The Supreme Court has shied away from addressing the constitutional status of private police. Outside the area of criminal investigations, the Supreme Court has developed an elaborate, notoriously muddled doctrine of state action; rigorously applied, that doctrine proves unable to distinguish private police convincingly either from public law enforcement or from the private citizenry. For their part, lower courts by and large have resolved the matter by resort to symbols: private security personnel are treated as private persons unless they have been officially deputized—regardless what additional legal powers, if any, that formal process confers.

This distinction is arbitrary, and it creates some conspicuous incongruities. But I will suggest that it makes rough sense. I will also argue that drawing a principled boundary between public and private policing is less important than recognizing the boundary as contingent and continually questioning whether rules found on one side could profitably be transported to the other. And I will suggest, ultimately, that the most urgent question raised by police privatization may not be whether to treat private security guards as state actors, nor even what substantive rules should govern the private police, but whether the state has some duty to provide everyone, rich or poor, with minimally adequate police protection.

A. The State Action Doctrine in Criminal Procedure

Perhaps the most basic and invariable principle of criminal procedure is that constitutional restrictions on policing—the limitations imposed by the Fourth, Fifth, and Sixth Amendments, the prophylactic rules of evidentiary exclusion constructed to
reinforce those limitations, and the analogous rules of state constitutional law—apply only to investigative action attributable to the government. As regards the Fourth Amendment, this principle received its definitive articulation more than seventy-five years ago, in *Burdeau v. McDowell*. Suspecting McDowell of fraud, his employer fired him and, allegedly without right, searched his office. The search turned up incriminating documents. McDowell’s employer handed over the documents to federal prosecutors. McDowell sued to get the documents back and for an order barring their presentation to the grand jury. The district court ruled in his favor, but the Supreme Court reversed.

Notwithstanding the apparent unlawfulness of the search and seizure, the Court found the Fourth Amendment inapplicable because the government had played no role. Writing for the majority, Justice William Day reasoned that the “origin and history” of the amendment “clearly show that it was intended as a restraint upon the activities of sovereign authority, and was not intended to be a limitation upon parties other than governmental agencies.” Accordingly, it was “manifest” that McDowell could complain of no violation of the Fourth Amendment because “no official of the Federal Government had anything to do with the wrongful seizure,” and “whatever wrong was done was the act of individuals in taking the property of another.” Furthermore, the Court saw “no reason” why that wrong, by individuals “unconnected with the government,” should prevent use of the documents in McDowell’s prosecution.

The Supreme Court has never revisited *Burdeau*; instead, it has treated both parts of the holding—the inapplicability of the Fourth Amendment and the inapplicability of the exclusionary rule—as obviously correct. Thus the Court has declared that “[i]t is well established, of course, that the exclusionary rule, as a deterrent sanction, is not applicable where a private party or a foreign government commits the offending act.” Similarly the Court has explained that it has “consistently construed” the Fourth Amendment “as proscribing only governmental action; it is wholly inapplicable ‘to a search or seizure, even an unreasonable one, effected by a private individual not acting as an agent of the Government or with the participation or knowledge of any governmental official.’”

The Court has interpreted the Fifth Amendment in a parallel fashion, although an unambiguous statement of this interpretation was longer in coming than *Burdeau*. At the close of the nineteenth century the Justices suggested that the Fifth Amendment barred the introduction against a federal defendant of any statement coerced from him, no matter by whom. More recently, however, the Court has described “the sole concern of the Fifth Amendment” as “governmental coercion” and has made clear that “coercive police activity is a necessary predicate” to a finding of involuntariness under either the Fifth Amendment or the Due Process Clause of the Fourteenth Amendment. Accordingly, lower courts without exception have refused to impose the prophylactic protections of *Miranda* on private interrogators.

Similarly, although the Supreme Court has never explicitly addressed whether the Sixth Amendment protection against uncounseled interrogation can apply to interrogations by private parties without government encouragement, the question has not bedeviled the lower courts, which have uniformly held the amendment inapplicable. A contrary answer
would be difficult to square with the Supreme Court’s repeated suggestion that the amendment prohibits only action by “the police and their informant” that is “designed deliberately to elicit incriminating remarks.” So, too, the Due Process Clauses prohibit prosecutions based on “outrageous” investigative techniques, but only when they are employed by the government. Virtually without exception, state constitutional restrictions on criminal investigations are similarly limited. Even the entrapment doctrine, which both federal and state courts describe as statutory rather than constitutional in origin, protects suspects only against instigation by government agents.

Although the state action doctrine has never received high marks for clarity, even its fiercest critics rarely complain about its application in the field of criminal procedure. When it comes to constitutional protections against overzealous policing, the state action doctrine has received something of a free pass. In part, no doubt, this dispensation is simply one manifestation of the general neglect of criminal procedure by constitutional scholars. But the state action doctrine does have some heightened intuitive appeal in the area of criminal procedure. Many of the most controversial state action cases have centered around the Equal Protection Clause of the Fourteenth Amendment, a provision which can easily be read—and may well have been intended—to require the government to redress certain abusive exercises of private power. In contrast, the Bill of Rights, and particularly the Fourth Amendment, seem more obviously and exclusively aimed at reining in the government. The libertarian thrust of these provisions bolsters the case for limiting their scope to abusive conduct by government agents.

So does the kind of conduct they prohibit. The most damaging private discrimination often has been carried out by institutions that are highly visible, easily regulated, and tightly affiliated with the government. Applying the Equal Protection Clause to these institutions makes a good deal of intuitive sense, both because these institutions often seem de facto arms of the state, and because, in any event, it seems reasonable to ask the government to ensure they act fairly. But the unofficial “searches” and “seizures” of greatest concern to most people are probably those carried out by criminals. It is one thing to assert that the Boy Scouts or the Elks are the state; it seems more of a stretch to say that of the Genovese crime family or the Crips. Calling criminals “state actors” seems silly. It may not be so silly to claim that the government has some obligation to protect people against crime, a higher obligation even than fighting private discrimination. Some people have read the Equal Protection Clause to impose just such an obligation, and I will suggest later that police privatization should prompt us to give that reading new consideration. But if affirmative duties of this kind are to be introduced into constitutional law, it seems more straightforward to use the Equal Protection Clause, not the back door of the Fourth Amendment.

There are good reasons, then, for the continued vitality of the state action doctrine in criminal procedure. The problem is how to maintain that doctrine when private policing increasingly blurs the line between public and private. One natural approach is to emphasize the distinctive risks posed by the police, and the distinctive costs to applying constitutional rights to private investigative activity. We might expect courts to probe the extent to which private policing raises the same risks as public policing, and the extent to
which constitutional regulation of the private police would entail the same costs as constitutional regulation of concerned neighbors and vigilant storekeepers. In fact, as we will see, such inquiries are vanishingly rare, and the reasons why are instructive.

B. State Action and Private Policing

1. The Supreme Court and Private Police

Although the Supreme Court has made clear through its unwavering adherence to Burdeau that the rules of the Fourth Amendment, and by implication the rest of constitutional criminal procedure, apply only to law enforcement activity “connected with” the government, it has provided remarkably little guidance regarding when if ever private policing should be deemed to satisfy that requirement.

The Court has directly addressed the state action problem in the context of private policing only once, in Griffin v. Maryland. Although Griffin was a criminal case, the question it posed was not of criminal procedure. The question was whether the Equal Protection Clause of the Fourteenth Amendment applied to the actions of a security guard employed by a private amusement park, but “deputized” by the county sheriff. The defendants were five black students who in 1960 visited the Glen Echo Amusement Park, flouting the park’s unwritten policy excluding blacks. Acting on the instructions of the park manager, security guard Francis Collins asked the students to leave and, when they refused, arrested them for trespassing. It developed that Collins was employed by a private firm, the National Detective Agency, under contract with the amusement park. It also developed that he had been deputized at the request of park management pursuant to a county ordinance authorizing the appointment of private-paid security personnel as “special deputy sheriffs”; this appointment gave him “the same power and authority as deputy sheriffs” within the boundaries of the park.

The students claimed their arrest and subsequent conviction constituted racially discriminatory state action in violation of the Equal Protection Clause. The Maryland Court of Appeals rejected this claim, finding no state action. The court noted that at the time of the arrests Collins “had on the uniform of the agency,” and that the arrests were carried out to further a private policy of segregation. His deputization, the court reasoned, “did not alter his status as an agent or employee of the operator of the park,” and under Maryland law that status alone permitted him to make a citizen’s arrest for any misdemeanor carried out in his presence. But the court also suggested that the capacity under which Collins acted was ultimately immaterial: whether he acted as a private citizen or a public officer, the court reasoned that the state was no more implicated in the park’s policy of segregation than it would be if the park had called “a regular police officer” to arrest the defendants for trespassing.

The Supreme Court reversed. Writing for the majority, Chief Justice Warren concluded, first, that Collins qualified as a state actor because he had “purported to exercise the authority of a deputy sheriff. He wore a sheriff’s badge and consistently identified himself as a deputy sheriff rather than as an employee of the park.” Second, Warren reasoned that the state action in this case violated the Equal Protection Clause because it
amounted to public enforcement of a private policy of racial segregation. He analogized it to a public agency enforcing racially discriminatory provisions of a private trust, a practice the Court had deemed unconstitutional seven years earlier.

*Griffin* was one of a series of postwar decisions loosening the state action limitation to facilitate an assault on private discrimination, cases in which the Court may well have been less concerned with doctrinal niceties than with the bottom line. Nonetheless the reasoning in *Griffin* merits close inspection, partly because the Supreme Court has said so little else about the constitutional status of private policing, and partly because there is a kind of submerged formalism to *Griffin* that foreshadows the lower court decisions we will survey later.

The *Griffin* majority rejected the argument that Collins had been “simply enforcing the park management’s desire to exclude designated individuals from the premises”—not because such a practice would be unconstitutional whenever the “designated individuals” were selected on the basis of race, but because Collins had been specifically “instructed . . . to enforce the park’s policy of racial segregation.” Ironically, the discrimination in *Griffin* therefore may have constituted state action precisely because the park enforced its policy through a privately paid security guard rather than by calling the police. Justice Clark’s concurring opinion made this explicit:

> If Collins had not been a police officer, if he had ordered the appellants off the premises and filed the charges of criminal trespass, and if then, for the first time, the police had come on the scene to serve a warrant issued in due course by a magistrate, based upon the charges filed, that might be a different case.

*Griffin* is thus a paradoxical decision: on the surface it appears a blow against formalism, but deeper down things are not so clear. The state’s complicity in the amusement park’s policy of segregation seems dependent, in the Court’s view, upon the merging of the roles of landowner’s agent and police officer in one individual. Moreover, although the Court rejected the notion that law enforcement is state action only when carried out by public employees, it strongly suggested that Collins qualified as a state actor only because he held himself out as one: the result might have been different, the majority implied, if Collins had not worn a badge, and—more importantly—had relied explicitly on his power of citizen’s arrest. Why these matters of form should be of such consequence was left unexplored, as was the question whether the analysis would differ had the defendants’ claims arisen under the Fourth, Fifth, or Sixth Amendment instead of under the Equal Protection Clause.

Any broad guidance *Griffin* could provide, moreover, was disavowed by the Court a decade later in *Flagg Bros. v. Brooks*. That case involved whether a self-help remedy granted to certain merchants by New York law constituted state action and thus was subject to the Due Process Clause of the Fourteenth Amendment; the Court answered in the negative. Dissenting from that result, Justice Stevens suggested in passing that “it is clear that the maintenance of a police force is a unique sovereign function, and the delegation of police power to a private party will entail state action.” In support he cited
Griffin. Writing for the majority, Justice Rehnquist “express[ed] no view” regarding whether and to what extent a city or state could “avoid the strictures of the Fourteenth Amendment” by delegating police functions to private parties, but he explicitly rejected Stevens’ reading of Griffin:

Contrary to Mr. Justice Stevens’ suggestion, this Court has never considered the private exercise of traditional police functions. In Griffin v. Maryland, the State contended that the deputy sheriff in question had acted only as a private security employee, but this Court specifically found that he “purported to exercise the authority of a deputy sheriff.” Griffin thus sheds no light on the constitutional status of private police forces, and we express no opinion here.

Nor has the Court expressed any opinion in the two decades following Flagg Bros.

2. Private Police in the Lower Courts

Unlike the Supreme Court, lower courts have been forced to address the constitutional status of private police. Like the Supreme Court, however, lower courts generally have avoided careful examination of the ways in which private police differ from their public counterparts and from ordinary citizens. Instead, less out of principle than out of habit, state and federal courts alike have simply declared the rules of constitutional criminal procedure inapplicable to private security personnel. Overwhelmingly, the lower courts have refused to treat private police as state actors, without ever offering a convincing explanation why. Private police have been held exempt from the Fourth Amendment and the Miranda rules—as well as from restrictions on entrapment and statutory disclosure requirements—all on the basis that they are private actors rather than government agents. Three representative cases will suffice to convey the tenor of these decisions.

- Lynn Johnson, a plain-clothes security guard at a Washington, D.C. department store, followed a shopper, Adelaide Lima, to a fitting room. Peering through louvers in the door, Johnson—according to her own later testimony, credited by the trial judge—saw Lima remove tags from a blouse and place it in her purse. Lima then left the dressing room, and Johnson again followed her. When Lima left the store, Johnson confronted her, identified herself as a store detective, physically restrained Lima, and escorted her to the store’s security office. At the office Lima was searched and the blouse was recovered. The trial judge suppressed the blouse in Lima’s subsequent prosecution for petit larceny, but the District of Columbia Court of Appeals reversed, holding the Fourth Amendment inapplicable to the actions of “mere employees performing security duties.”

In support of its holding, the court stressed that “[s]earches and seizures by private security employees have traditionally been viewed as those of a private citizen and consequently not subject to Fourth Amendment proscriptions.” The court distinguished in this regard a “licensed security officer,” such as Johnson, from “commissioned or deputized special police officers,” who although privately employed are “commonly vested by the state with powers beyond that of an
ordinary citizen.” A licensed security officer, by contrast, “has only the power of arrest of an ordinary citizen.” Far from “equat[ing] the powers of the security guard with the powers of police officers,” the licensing statute made “a conscious effort to distinguish the security guard from special policemen or regular policemen by prohibiting the wearing of uniforms or badges resembling those officers.”

Lima argued “that security employees who go around ‘walking, talking, acting and getting paid like policemen’ should in fact be treated as policemen for the purposes of the Fourth Amendment.” But the court of appeals expressly rejected the suggestion “that application of the Fourth Amendment can be resolved by looking at the nature of the activities performed by security employees.” Like other private individuals, the court observed, private police are under no public duty to make arrests; they do so purely out of “[s]elf interest.” The court further reasoned that “[s]olely permitting the act is insufficient ‘to support a finding of state action,’” and “[t]he fact that the private sector may do for its own benefit what the state may also do for the public benefit does not implicate the state in private activity.”

- Nicholas Antonelli, a New York dockworker, was driving off his worksite when he was stopped by a security guard. At the guard’s request, Antonelli opened the trunk of his car. Inside were six burlap bags filled with several thousand dollars’ worth of stolen clothing. Antonelli proceeded to make incriminating statements to the guard. He later sought to suppress these statements on the ground that the guard had not advised him of his *Miranda* rights. His motion was denied, he was convicted, and the Second Circuit affirmed.

The court assumed arguendo that Antonelli’s statements resulted from “custodial interrogation”—that is, from the kind of actions that would trigger *Miranda* if carried out by public law enforcement officers. But the court noted that the security guard “had no pertinent official or de facto connection with any public law enforcement agency,” was employed “to protect the private property on the docks,” and had not been influenced, instigated, or assisted in his dealings with Antonelli by any government officer. *Miranda*, the Second Circuit reasoned, “was concerned solely with activity by the police or other . . . government agencies,” and “[a] private security guard stands no differently from the private citizen who has employed him.”

The court opined that “[i]t would be a strange doctrine that would so condition the privilege of a citizen to question another whom he suspects of stealing his property that incriminating answers would be excluded as evidence in a criminal trial unless the citizen had warned the marauder that he need not answer.” This would be particularly strange, in the court’s view, because “[t]he federal exclusionary rule enforcing adherence to the intendment of the Fifth Amendment, like the Fourth Amendment, has long been construed as ‘a restraint upon the activities of sovereign authority’ and not as ‘a limitation upon other than
government agencies.’” Finally, the court pointed out that state courts addressing the question had held *Miranda* inapplicable to questioning by private parties.

- Adam Taylor was sitting with three acquaintances under the Santa Cruz Beach Boardwalk when he was approached by two uniformed guards employed by Seaside Land Company, which owned the boardwalk. One of Taylor’s companions was smoking a marijuana cigarette, and Taylor had on his lap a baggie containing something green. One of the guards, Officer Kerr, asked Taylor for the baggie and for identification. Taylor gave Kerr the baggie and said he had no identification. Kerr asked Taylor if he had any more drugs on him and if he would consent to a search. Taylor denied having drugs and agreed to a search, although exactly what kind of search is not completely clear. According to Taylor, he consented only to a search of his fannypack, which he opened and handed to Kerr, and in which Kerr found no contraband. Kerr, however, proceeded to pat down Taylor’s pants, and in this manner discovered four more baggies containing marijuana and two baggies containing LSD. Kerr then handcuffed Taylor, took him to a detention center operated by Seaside Land Company, and called the police.

Following denial of his suppression motion, Taylor entered a conditional guilty plea and then appealed. The court of appeal affirmed. The court found no need to determine the legality of Kerr’s search because it concluded that the exclusionary rule and the Fourth Amendment applied only to the police, and “searches and seizures by private security employees have traditionally been viewed as those of a private citizen and consequently not subject to Fourth Amendment proscriptions.” The function of carrying out arrests was not one “traditionally exclusive to the state” because “[t]here have been citizen arrests for as long as there have been public police—indeed much longer.”

Nor, despite allegations that boardwalk security guards shared facilities with the police and had access to police radios, was the appellate court convinced that the security guards “operate[d] jointly with the police,” so that “their action should be imputed to the state.” The court noted that “the impetus for the arrests came from the security guards,” and that there was no evidence of any “contracts or agreements” between the security force and the police. Furthermore, and perhaps most basically, the appellate court stressed that the security guards simply were not the police:

> [T]here is no evidence . . . that the security guards’ uniform was in any way similar in color to that of the Santa Cruz police department. The guards wore badges and shoulder patches marked “security.” They did not carry guns. Although they did not identify themselves verbally as security guards, they did not verbally claim to be police officers. The evidence showed that their private purpose was to patrol Seaside Land Company’s property.
Summing up, the court quoted an instructional pamphlet issued by state regulators: “A security guard is not a police officer. Guards do not have the same job duties as police officers; they do not have the same training; and they do not have the same powers according to law.”

The decisions in *United States v. Lima*, *United States v. Antonelli*, and *People v. Taylor* are illustrative not only because, in each case, private security guards were treated as private rather than public actors, but also because the analysis in each case consisted largely of an appeal to precedent—how security guards have traditionally been viewed—coupled with the observation that private guards are distinguishable from public law enforcement officers—in their duties, in their powers, and in their dress. When private guards are more difficult to distinguish from the public police—either because the guards are in fact off-duty police officers, or because they have been sworn in as special patrolmen with police powers not shared by ordinary citizens—courts often treat the guards as agents of the state. But so long as private police personnel occupy an awkward middle ground between public law enforcement and private citizens, courts overwhelmingly focus on the distinctions between the guards and the public police. Rare are the decisions questioning whether those distinctions should make a difference.

Partly because such decisions are so rare, they have a difficult time remaining good law. In the late 1970s and early 1980s, the highest courts of California and Montana, alarmed by the privatization of law enforcement, warmed to the proposition that private police, even without special powers, should be treated as state actors. Writing in 1981, Professor Burkoff could express optimism that state courts, interpreting state constitutions, might lead the way toward a broad “[r]econsideration of the *Burdeau* private search doctrine.” But the moment passed, and the renegade states returned to the fold—a return triggered in California by popular initiative, and in Montana by further judicial contemplation of the vast weight of contrary case law. In 1987, the increasing prominence of private security forces prompted West Virginia’s highest court to hold state constitutional restrictions on searches, seizures, and custodial interrogations applicable to all statutorily authorized detentions, whether carried out by public officers or private citizens. That decision, too, now appears a dead letter.

Less out of principle than out of habit, state and federal courts alike resolve the ambiguous status of private police by grouping them with private citizens and distinguishing them sharply from the public police. Despite occasional rhetoric to the contrary, the decisions are conspicuously formalist; virtually everything turns on whether the state has vested private personnel with an official title. The enormous weight these cases place on symbolism and precedent, coupled with the Supreme Court’s silence regarding the constitutional status of private police, suggests this is an area ripe for rigorous doctrinal analysis.

As we will see, however, the suggestion is overoptimistic. State action doctrine—famously dismissed as a “conceptual disaster area,” and more charitably described by the Court itself as “not . . . a model of consistency”—is all but useless in determining the proper treatment of private security personnel. But while state action doctrine has little to
tell us about the private police, the private police have lessons to teach us about state action doctrine, and particularly about its application in the field of criminal procedure. It is therefore worth our time to try applying the general doctrines of state action to the private police.

3. The “Principles” of State Action

As good a starting point as any is \textit{Lugar v. Edmondson Oil Co.}, which has served as the touchstone for most of the Supreme Court’s recent discussions of state action. The actual question in \textit{Lugar} was not constitutional but statutory: whether a private creditor’s prejudgment attachment of a debtor’s property constituted conduct “under color of” state law within the meaning of 42 U.S.C. § 1983. The Court reasoned, however, that “the state-action and the under-color-of-state-law requirements are obviously related,” and, more particularly, that any conduct that qualifies as state action necessarily is also action under color of state law.

Writing for the Court, Justice White then set forth a two-part test for determining whether challenged conduct is “fairly attributable” to the government:

First, the deprivation must be caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the State or by a person for whom the State is responsible. . . Second, the party charged with the deprivation must be a person who may fairly be said to be a state actor. This may be because he is a state official, because he has acted together with or has obtained significant aid from state officials, or because his conduct is otherwise chargeable to the State.

More recently, the Court has described the second inquiry as “often . . . factbound,” but guided by “certain principles of general application.” Specifically,

[our precedents establish that, in determining whether a particular action or course of conduct is governmental in character, it is relevant to examine the following: the extent to which the actor relies on governmental assistance and benefits, whether the actor is performing a traditional governmental function, and whether the injury caused is aggravated in a unique way by the incidents of governmental authority.

The source for this list of relevant “principles,” \textit{Edmonson v. Leesville Concrete Co.}, is one of two recent cases in which the Court has applied the \textit{Lugar} “framework” to the race-based use of peremptory challenges. These cases merit a brief detour, not only because they clarify \textit{Lugar} but also because, like most criminal procedure cases, they involve claims raised in the context of trial proceedings.

In \textit{Edmonson}, the Court held that a private litigant’s exercise of peremptory strikes based on the race of potential jurors constituted state action in violation of the Equal Protection Clause of the Fourteenth Amendment; subsequently, in \textit{Georgia v. McCollum}, the Court
reached the same conclusion about race-based peremptory challenges by a criminal defendant. In both cases, the Court thought it beyond question that peremptory strikes satisfy the first *Lugar* requirement because they “are permitted only when the government, by statute or decisional law, deems it appropriate to allow parties to exclude a given number of persons who otherwise would satisfy the requirements for service on the petit jury.” Turning to the second *Lugar* requirement, both *Edmonson* and *McCollum* found that all three “principles” supported a finding of state action. First, civil litigants and criminal defendants exercising peremptory strikes rely heavily on “government assistance and benefits” because “the peremptory challenge system, as well as the jury system as a whole, ‘simply could not exist’ without the ‘overt, significant participation of the government.’” Second, peremptory challenges involve performance of a “traditional governmental function” because “the objective of jury selection proceedings is to determine representation on a governmental body.” Third, the harm produced by race-based challenges to jurors is “aggravated in a unique way by the incidents of governmental authority”: “the racial insult” inherent in discrimination is “made more severe because the government permits it to occur within the courthouse itself.”

To a reader uninitiated in the mysteries of state action jurisprudence, the reasoning of *Edmonson* and *McCollum* might seem oddly roundabout. Why bother searching for state action in the challenge raised by the criminal defendant or private civil litigant, when state action is so manifest in the trial court’s removal of the challenged juror? After all, race-based challenges affect the composition of juries only because trial judges go along with them. By straining to find state action in the litigant’s request rather than in the judge’s compliance, the Court appeared to complicate things needlessly. Led by Justice O’Connor, three dissenters in *Edmonson* argued plausibly that “attorneys represent their clients,” and that “[i]t is antithetical to the nature of our adversarial process . . . to say that a private attorney acting on behalf of a private client represents the government for constitutional purposes.” One of those three, Justice Scalia, separately characterized *McCollum* as reducing *Edmonson* “to the terminally absurd: [a] criminal defendant, in the process of defending himself against the state, is held to be acting on behalf of the state.” But even the *Edmonson* dissenters tacitly conceded that a peremptory challenge becomes effective only when the trial judge “acquiesce[s] . . . by excusing the juror.” Why not simply declare the acquiescence unconstitutional and avoid the whole problem?

The difficulty, of course, is that when trial judges acquiesce in peremptory challenges, they act with at least facial neutrality, and the Court has held that facial neutrality—that is, the absence of purposeful discrimination—generally suffices to satisfy the Equal Protection Clause. If the only state action in *Edmonson* and *McCollum* was what the trial courts did, then the Constitution was violated only if the Equal Protection Clause prohibits the government not only from *making* discriminatory choices, but also from *honoring*—that is, giving legal effect to—discriminatory choices made by others. Delegating dirty work to private parties was the classic tactic of Jim Crow, and a tactic the Court was loath to countenance. But the central lesson of legal realism is that all legal rules giving effect to private decisions amount to a delegation of state authority. As a consequence, if the Constitution prohibits the government from acquiescing in all private conduct in which the government could not itself engage, it goes far toward prohibiting
private parties from doing what the government cannot, and precious little is left of the
state action doctrine. This last result might not leave all observers dissatisfied, but it is
one the Court has never been willing to accept.

Among the lessons of Edmonson and McCollum is that even in cases in which some kind
of state action appears obvious—and that may, after all, include all cases—application of
constitutional norms may turn on who counts as part of the state. With this background in
mind, let us consider the private police in light of the three factors the Court identified in
Edmonson and McCollum as relevant to a determination whether particular conduct
should be deemed “governmental in character”: the degree of reliance on “governmental
assistance and benefits,” the presence or absence of a “traditional governmental
function,” and the extent to which the harm is exacerbated in a “unique” fashion by
“incidents of governmental authority.” Each of these factors, we will see, suggests if
taken at face value that private security guards should typically, or at least frequently, be
categorized as state actors; the problem is that none of them provides a convincing basis
for distinguishing private security guards from other private individuals.

a. Governmental Assistance and Benefits

The Court in Edmonson cited two cases illustrating the relevance, for state action
purposes, of “the extent to which the actor relies on governmental assistance and
benefits”: Burton v. Wilmington Parking Authority and Tulsa Professional Collection
Services, Inc. v. Pope. In Burton, the Court concluded that a restaurant’s refusal to serve
blacks was state action, and hence covered by the Equal Protection Clause, because the
restaurant leased space in a parking facility owned and operated by the State of Delaware,
placing the restaurant and the state in “a position of interdependence.” Pope found state
action in the operation of an Oklahoma statute requiring creditors to file claims with an
estate within two months of public notice of the start of probate proceedings. The Court
was quick to reaffirm that “[p]rivate use of state-sanctioned private remedies or
procedures does not rise to the level of state action,” but reasoned that the involvement of
Oklahoma probate courts in the operation of the two-month time bar was “so pervasive
and substantial” as to constitute state action: a probate court appointed the executor or
executrix, who then triggered the time bar by publishing notice to creditors; the court
appeared routinely to order the executor or executrix to provide such notice; and
Oklahoma law required proof of the notice to be filed with the probate court.

Private security firms typically are not tenants in publicly owned buildings and do not
work in close cooperation with the courts. But as Professor Burkoff among others has
pointed out, they enjoy other, arguably comparable forms of government assistance and
benefits, even when they are not hired by the government. Cooperative arrangements
between public and private police were already common in the 1970s, and since then they
have grown more common and more extensive. The arrangements vary but usually
involve, at a minimum, coordination of strategies and sharing of information, often
including arrest and conviction records. Some police departments conduct training
programs for private security personnel; others share facilities and equipment. And
because so many police officers moonlight as security guards, “[p]rivate policing uses
police that have been recruited, trained, and supported by government.” Finally, in addition to these forms of tangible assistance, private security firms increasingly gain credibility and legitimacy by working in cooperation with the public police. In return, the police gain the opportunity to leverage their own resources—either by out-contracting or simply by allowing private security to dampen the demand for public policing. The road thus is straight and short from Burton and Pope to a conclusion that much if not all private policing constitutes state action.

The problem with this road is that it goes too many places. Public law enforcement officials cooperate with and assist all kinds of entities and individuals, not just private security firms. They work with banks to deter robberies, help train merchants to cut shoplifting, and consult with community groups on plans to make streets and parks safer. All these forms of cooperation, moreover, are sharply increasing: part of the point of “community policing” is to make “[m]obilizing communities for their own defense . . . a central strategy of policing, a mainline function that is the responsibility of every officer.” If it is difficult to distinguish the government benefits enjoyed by private security firms from the assistance deemed dispositive in Burton and Pope, it is at least as hard to differentiate those benefits from the similar support the police provide to the broader community.

Nonetheless one’s instinct is that the private police are different from banks, merchants, and community groups. Much of this instinct, of course, stems from the fact that the private police, unlike banks, merchants, and community groups, bear a particularly close resemblance to familiar government agents. They look and act like cops. So one turns naturally to the second of the Edmonson-McCollum factors: the presence of a “traditional governmental function.”

b. Traditional Governmental Function

The idea that private parties become state actors when they step into the shoes of the government—the so-called “public function” doctrine—is traditionally associated with the Supreme Court’s decisions in Marsh v. Alabama and Terry v. Adams, both of which were cited in Edmonson and McCollum. Marsh and Terry, decided half a century ago, are explosive in their potential but have been remarkably limited in their actual consequences. Indeed, few Supreme Court cases have been cited by litigants so frequently with such limited success. Litigants seeking to subject private security personnel to constitutional constraints have been particularly fond of Marsh. To see why that decision seems to promise them so much but has delivered so little, we need to examine, in some detail, both the case itself and the Supreme Court’s subsequent efforts to make sense of it.

Like Griffin, Marsh was a trespassing prosecution. The case arose in Chickasaw, Alabama, a company town owned by the Gulf Shipbuilding Corporation. Grace Marsh, a Jehovah’s Witness, was arrested—by a deputy sheriff paid by Gulf Shipbuilding—when she distributed religious literature on a sidewalk in Chickasaw, in violation of a company rule against unauthorized solicitation. She challenged her conviction on First Amendment
grounds, but the Alabama courts reasoned that Gulf Shipbuilding, as a private party, was not governed by the First Amendment, or by the Due Process Clause of the Fourteenth Amendment, through which the First Amendment applies to the states.

The Supreme Court reversed. Writing for the Court, Justice Black observed that Chickasaw had “all the characteristics of any other American town,” except that “all the property interests in the town are held by a single company.” Given the importance of free speech, the Court found this difference immaterial: “[w]hether a corporation or a municipality owns or possesses the town the public in either case has an identical interest in the functioning of the community in such manner that the channels of communication remain free.” And the Court found the appeal to property rights unpersuasive in this context, reasoning that “[t]he more an owner, for his advantage, opens up his property for use by the public in general, the more do his rights become circumscribed by the statutory and constitutional rights of those who use it.”

The phrase “public function” appeared in *Marsh* not to describe the streets and sidewalks of Chickasaw, but in reference to a comparison the Court drew to privately owned highways, bridges, ferries, and railroads. Because operation of these facilities “is essentially a public function,” the Court noted, it was “subject to state regulation,” and the Court reasoned that a state could neither require nor allow such regulated facilities to operate in a manner that discriminated against interstate commerce. Justice Black conceded this situation was “not directly analogous” to the case at hand, and Justice Frankfurter, concurring separately, specifically declined to join this portion of the majority opinion: “It does not seem to me to further Constitutional analysis to seek help for the solution of the delicate problems arising under the First Amendment from the very different order of problems which the Commerce Clause presents.”

The Court’s reasoning in *Marsh* has never been easy to pin down. In his perceptive analysis of the case, Julian Eule identified “strands of no fewer than four theories—public function, appearance, waiver, and government complicity”—and hints of a fifth: “a suggestion of a guaranteed right to be informed and a concomitant government responsibility to ensure the free flow of information.” Over the past five decades, however, *Marsh* has come to be associated most closely with the idea that, at least in some cases, constitutional constraints should apply to private parties who carry out “public functions”—or, in the language of *Edmonson* and *McCollum*, “traditional governmental functions.” Identifying such functions, however, has been difficult. *Marsh* itself has been all but limited to its facts: Justice Black later characterized the holding as limited to private property that had “all the attributes of a town,” and, after some vacillation, the Court echoed that view. This may be why the Court today, when referring to the public function doctrine, commonly couples its reference to *Marsh* with a citation to *Terry*.

At issue in *Terry* were the activities of a private organization, the Jaybird Democratic Association, which automatically enrolled as members all white voters in Fort Bend County, Texas and excluded blacks. For more than sixty years, the association had conducted elections to select candidates to run for nomination for county offices in the
Democratic primary, and the winners of the Jaybird elections almost invariably had run unopposed in both the Democratic primaries and the general elections. The Court concluded this system violated the Fifteenth Amendment guarantee against racial abridgements of the right to vote, but was unable to produce a majority opinion identifying why constitutional limitations applied in the first place. A plurality, led by Justice Clark, found state action in the Jaybird election, concluding that the association operated, with the state’s consent, as “the decisive power in the county’s recognized electoral process.” Concurring separately, Justice Frankfurter reasoned that the Jaybird elections themselves were beyond constitutional purview, but that Texas had violated the Fifteenth Amendment by allowing the results of those elections to become, “by virtue of the participation and acquiescence of State authorities,” the de facto results of the Democratic primaries. The remaining three Justices concurring in the result nicely sidestepped the issue by locating the constitutional violation in “[t]he effect of the whole procedure, Jaybird primary plus Democratic primary plus general election.”

Given this ambiguity at the outset, it should come as no surprise that, as a state action decision, *Terry* has been little more generative than *Marsh*. The Supreme Court has relied significantly on the public function doctrine in only one subsequent case, *Evans v. Newton*, which concerned a park operated for decades by the city of Macon, Georgia, but returned to private hands, under the terms of its original devise, when the city concluded it could not constitutionally continue to exclude blacks. Writing for the Court in *Evans*, Justice Douglas read *Marsh* and *Terry* to require imposition of constitutional constraints on any individuals or groups the state allowed to exercise “powers or functions governmental in nature,” and he noted—tantalizingly for our present purposes—that a park, unlike a golf club or social center, “is more like a fire department or police department that traditionally serves the community.” In language bristling with implications for private policing in the 1990s, Justice Douglas reasoned that “[m]ass recreation through the use of parks is plainly in the public domain,” and therefore that “state courts that aid private parties to perform that public function on a segregated basis implicate the State in conduct proscribed by the Fourteenth Amendment.”

But what Justice Douglas gave with one hand he took away with the other, stressing that this particular park long had been, and apparently still was, “an integral part of the City of Macon’s activities” being “swept, manicured, watered, patrolled, and maintained by the city.” Macon thus remained “entwined” in the park’s management and control, and Justice Douglas made clear, in the key sentence of his opinion, that the Court held only “that where the tradition of municipal control had become firmly established, we cannot take judicial notice that the mere substitution of trustees instantly transferred this park from the public to the private sector.”

*Evans* remains for all that the high point of the public function doctrine; today the doctrine holds far less promise for litigants hoping to apply constitutional restrictions to nominally private action. The pivotal move came in *Jackson v. Metropolitan Edison Co.* Writing for the Court in that case, Justice Rehnquist reconceived the doctrine as applying only to the exercise of powers that “traditionally” have been “exclusively reserved to the State.” Since *Jackson*, no functions other than conducting elections for public office and
running an entire town have been deemed to qualify. This has not troubled the Court. Writing again for the majority in *Flagg Bros. v. Brooks*, Justice Rehnquist noted that, “[w]hile many functions have been traditionally performed by governments, very few have been ‘exclusively reserved to the State.’” Resolution of debt disputes, *Flagg Bros.* held, falls outside the category of exclusive public functions because “[c]reditors and debtors have had available to them historically a far wider number of choices than has one who would be elected public official, or a member of Jehovah’s Witnesses who wished to distribute literature in Chickasaw, Ala., at the time *Marsh* was decided.”

In theory, policing may be different. The Court made a point to leave this question open in *Flagg Bros.*:

> [W]e would be remiss if we did not note that there are a number of state and municipal functions not covered by our election cases or governed by the reasoning of *Marsh* which have been administered with a greater degree of exclusivity by States and municipalities than has the function of so-called “dispute resolution.” Among these are such functions as education, fire and police protection, and tax collection. We express no view as to the extent, if any, to which a city or State might be free to delegate to private parties the performance of such functions and thereby avoid the strictures of the Fourteenth Amendment.

As I have discussed earlier, some common intuitions support the notion that policing is inherently public and governmental in a way that, say, resolution of debt disputes or even control of parklands is not. Far from being twentieth-century inventions, these intuitions have ancient roots in the concept of the king’s peace—the original public function. And they provide the basis for the most common argument for treating private police as state actors. Indeed, if policing is not a public function, it is hard to imagine that much else is. Policing, we might say, is a test case for the vitality of the public function doctrine.

But any effort to portray policing as a power “traditionally exclusively reserved to the State” encounters two imposing obstacles. The first is that, despite the Court’s suggestion in *Flagg Bros.*, no aspect of policing, neither patrol nor detection, has ever been “exclusively” performed by the government, and all have at one point or another been left largely to private initiative. Indeed, as we have seen, the history of public policing is virtually inseparable from the history of private policing. A strong argument can be made that during the past century *some* part of policing has become “exclusively reserved to the state”—perhaps the role of public law enforcement as the peacekeepers of last resort. But that role remains public even in Southern California, where police privatization has gone further than anywhere else in the country. The roles in which private firms are increasingly supplanting public law enforcement—notably, routine patrol—are indeed roles that governments have exercised in most of the United States since the nineteenth century, but they are difficult if not impossible to characterize as exclusively reserved to the public sector. If policing is a test case for the vitality of the public function doctrine, the doctrine can safely be declared moribund.
To some extent this problem can be traced to *Jackson* and *Flagg Bros.*: it arises because the Supreme Court has limited the public function doctrine to functions that traditionally have been exclusively public. Were the Supreme Court to retract that limitation, the difficulty would largely disappear. But then the Court would be faced with the forbidding task of finding some other basis for distinguishing functions that are essentially public from those that are essentially private. Not so long ago the Court abandoned this task as hopeless in the context of the Tenth Amendment, recognizing among other things that “the ‘traditional’ nature of a particular governmental function can be a matter of historical nearsightedness; today’s self-evidently ‘traditional’ function is often yesterday’s suspect innovation.”

No doctrinal change, moreover, could obviate a second, more fundamental obstacle to deeming private policing state action on the ground that it carries out a “traditional governmental function”: the problem of distinguishing private security personnel from other private individuals, like the agents of McDowell’s employer in *Burdeau*. As long as we have a state action doctrine in criminal procedure, a focus on public function ultimately does little better than a focus on government benefits and assistance in making a case for treating private security guards but not neighbors, shopkeepers, and flight attendants as state actors.

There is one difference, of course, between security guards on the one hand and flight attendants on the other: security guards, by design, look like cops. They wear similar uniforms and similar badges; often they drive similarly marked patrol cars. As Adelaide Lima argued unsuccessfully, they “go around ‘walking, talking, acting and getting paid like policemen.’” Concentrating on such matters of form may not be as silly as it seems. After all, Justice Black’s opinion for the Court in *Marsh* stressed among other things the physical resemblance between Chickasaw and “any other American town,” noting that the property owned by Gulf Shipbuilding could not be distinguished from the surrounding neighborhood “by anyone not familiar with the property lines.” And uniforms, far more than powers, were a focal point—often the focal point—of nineteenth-century debates over the formation of centralized, professional police departments. From a practical standpoint, as well, giving doctrinal significance to the appearance of security guards has some appeal: if private firms choose to “borrow the halo and symbols of authority (uniform, gun, badge, accessories) associated with public enforcement,” perhaps they should also assume the responsibilities imposed on the public police.

The difficulty with this argument is twofold. First, of course, there is the familiar question of degree: how much resemblance should be required before private police become state actors? The second, more serious problem has to do with agency. Particularly in the field of criminal procedure, the state action doctrine typically is less about what constraints shall be placed on private action, and more about when the state shall be held accountable for private action. Allowing this decision to turn on how private agencies choose to clothe themselves seems odd: why should such private choices determine whether, for example, evidence should be admissible in a criminal case brought by the state? One possible answer is that the private choices here are made possible by a public choice: by allowing private guards to dress like police officers, the
state in effect has delegated symbolic authority to private security firms. But this answer simply brings the difficulty into sharper relief because it suggests that the state “action” here is really inaction—acquiescence—and, for reasons already discussed, once state action is understood to include acquiescence, the state action doctrine begins to unravel.

c. Incidents of Governmental Authority

One encounters similar problems when assessing private policing in light of the last of the three “principles” that the Court in Edmonson and McCollum suggested should guide determinations whether to treat particular conduct as “governmental in character,” namely “whether the injury caused is aggravated in a unique way by the incidents of governmental authority.” As authority for this last principle, the Court cited “the most problematic and controversial of the state action cases,” Shelley v. Kraemer.

We have become so accustomed to thinking of Shelley as perplexing that it is easy to forget how straightforward the decision can appear at first reading. At issue in the case was whether judicial enforcement of racial restrictive covenants violated the Equal Protection Clause of the Fourteenth Amendment. Without dissent the Court held that it did. When judges enforce such agreements, the Court reasoned, they put “the full coercive power of government” and “the clear and unmistakable imprimatur of the State” at the disposal of those wishing to limit property ownership to whites; this goes beyond “merely abstain[ing] from action, leaving private individuals free to impose such discriminations as they see fit.” And for purposes of the Fourteenth Amendment, the Court explained, the phrase “[s]tate action . . . refers to exertions of state power in all forms,” including the actions “of judicial officers in their official capacities,” and specifically including the exercise by those officials of “the power of the State to create and enforce property interests.” Today, now that the lessons of legal realism have become commonplace, the reasoning in Shelley can seem unremarkable, if not “irresistibly correct.”

The problem, of course, is that state action similar to that involved in Shelley can be found whenever a court does almost anything, from enjoining a trespass, to awarding damages for an assault, to enforcing a contract. And if all the state action doctrine means is that our private conduct is free from constitutional constraint as long as we do not want government protection against trespasses, torts, or breaches of contract, it does not mean much. Recognizing this difficulty, the Supreme Court simply refused to take Shelley to its logical conclusion, and over time limited the decision for practical purposes to cases involving racially restrictive covenants. Writing for the Court recently, Justice Scalia summed up the now prevailing view of Shelley: “[a]ny argument driven to reliance upon an extension of that volatile case is obviously in serious trouble.”

Nonetheless the decision remains on the books, and it hardly has to be stretched to reach the case of a criminal prosecution based in part on evidence obtained by private police. The Fourth, Fifth, and Sixth Amendments are all enforced, for the most part, by excluding evidence in criminal cases. The grounds for exclusion differ, depending upon which constraint is at issue. Where the Fifth Amendment is concerned, the introduction
of compelled testimony at trial is understood to be the violation; the Fourth Amendment exclusionary rule, in contrast, has been justified primarily as a means of deterring violations, or avoiding judicial acquiescence and complicity in violations; whether the Sixth Amendment is more like the Fourth or the Fifth in this respect remains unclear. But regardless whether evidentiary exclusion is required because admission of the evidence is itself wrongful, or because admission of the evidence would constitute judicial encouragement of, or acquiescence or complicity in, some collateral wrongdoing, it is hard to avoid seeing state action in the use of particular evidence against a criminal defendant—no matter whether the evidence was generated by a government agent or a private individual.

As Professor Burkoff has argued,

[when the State affirmatively accepts illegally seized evidence in its criminal justice system, thereby authorizing or encouraging actions by private parties that would be unconstitutional if performed by governmental officials, it ignores reality to then assert that there is no “sufficiently close nexus between the State and the challenged action.”]

In an important respect the nexus is closer than in Edmonson, which dealt with peremptory challenges by private civil litigants, and McCollum, which concerned peremptory challenges by criminal defendants, because evidence against the accused in a criminal case not only is ruled admissible by a judge, but also is offered in the first instance by a prosecutor, “a quintessential state actor.”

But this really is not an argument for treating private police as state actors: it is an argument for overruling Burdeau and altogether eliminating the state action requirement in criminal procedure, at least when, as is usual, the remedy sought is suppression of evidence against an accused. Like the other two factors listed in Edmonson and McCollum as guides for identifying state action, the third factor—“whether the injury caused is aggravated in a unique way by the incidents of governmental authority”—thus provides no basis for distinguishing private police either from public police or from other private individuals.

4. Disaggregating the State Action Problem

Before turning to the lessons to be drawn from the remarkable uselessness of state action doctrine in cases involving the private police, one additional feature of that doctrine remains to be explored. In criminal procedure as elsewhere, state action doctrine is strikingly unified. Courts generally have treated as interchangeable the questions whether private security personnel are state actors for purposes of the Fourth Amendment exclusionary rule, the requirements of Miranda, and the color-of-law element of 42 U.S.C. § 1983. But perhaps the most common scholarly suggestion for disentangling state action doctrine is to pay greater attention to the precise nature of the right asserted and the remedy requested. One might hope, therefore, that the problem sketched in the preceding pages could be dissolved by disaggregation, replacing a single, abstract question with a range of more concrete inquiries. Unfortunately, disaggregation leaves
the dilemmas of state action doctrine intact, although it does allow us to see them more clearly.

Take, for example, a case like *Lima* or *Taylor*, presenting the question whether the exclusionary rule should apply to an allegedly illegal search by private police. There is no particular reason this question must be answered the same way as the separate question, posed in cases like *United States v. Antonelli*, whether the *Miranda* rules should apply to security guards. The concerns in Fourth Amendment cases may differ from those in *Miranda* cases, and private policing—or some forms of some private policing—may raise one set of concerns but not the other. Furthermore, a case like *Lima* or *Taylor* can be viewed as asking not one state action question, but two. The first is whether the Fourth Amendment provides a substantive right against unreasonable searches or seizures by private police; the second is whether the Fourth Amendment or any other part of the Constitution requires the suppression of evidence obtained improperly by private police.

The answer to the second question could of course depend on what makes the search or seizure improper; the Constitution might require an exclusionary remedy for certain kinds of transgressions but not for others. For example, it might require suppression when the Constitution is violated, but not when statutes are violated. This is in fact how the Supreme Court has interpreted the remedial requirements of the Fourth Amendment for searches and seizures by public law enforcement personnel. Ever since *Olmstead v. United States*, the Court consistently has found no constitutional right to suppression for nonconstitutional violations. But the Court has never satisfactorily explained this result, and neither has anyone else. One possible explanation is that the Constitution requires no particular remedy to secure rights it does not itself grant: if the government remains free to repeal the right, it should have the lesser power to decline enforcing the right through evidentiary exclusion. The problem with this explanation is that, here as elsewhere, it is not obvious that the greater power should include the lesser.

The Court itself seems to have recognized the problem. After concluding in *United States v. Caceres* that IRS regulations on the use of electronic surveillance in tax investigations were not constitutionally mandated, the Court noted that “it does not necessarily follow . . . as a matter of either logic or law, that the agency had no duty to obey them.” Nonetheless the Court declined on prudential grounds to require suppression of evidence the IRS had obtained in violation of its own rules: “we cannot ignore the possibility that a rigid application of an exclusionary rule to every regulatory violation could have a serious deterrent impact on the formulation of additional standards to govern prosecutorial and police procedures.” A similar argument could of course be made against applying the exclusionary rule to statutory violations: perhaps this would deter Congress and state legislatures from enacting further restrictions. The argument can even be—and has been—extended to constitutional violations: maybe “[j]udges do not like excluding bloody knives, so they distort doctrine, claiming the Fourth Amendment was not really violated.” No version of the argument is wholly implausible, but at this point only the last can claim to be anything more than pure speculation.
Just as exclusion could be warranted without a prior violation of the Fourth Amendment, so prior violation of the Fourth Amendment does not necessarily require exclusion. The Supreme Court has instructed that the exclusionary rule should apply only when its incremental value in deterring constitutional violations outweighs its costs for “truth-finding” and law enforcement. The Court has allowed the use, for example, of certain evidence “obtained in the reasonable good-faith belief that a search or seizure was in accord with the Fourth Amendment.” In similar spirit, lower courts sometimes reason that applying the exclusionary rule to the activities of private police would do no good. As Professor LaFave has expressed the point, “the exclusionary rule would not likely deter the private searcher, who is often motivated by reasons independent of a desire to secure criminal conviction and who seldom engages in searches upon a sufficiently regular basis to be affected by the exclusionary sanction.” But LaFave also concludes that

[w]here private police actually supplant the public police or deal regularly with the public, particularly if it may be said they are not disinterested in criminal convictions as an aid to the private objectives of their employer, it would be sound as a matter of law and policy to hold those police subject to the commands of the Fourth Amendment.

LaFave’s discussion is illuminating in two ways. First, LaFave recognizes, as most courts have not, that simple cost-benefit assessments cannot justify sharply limiting the exclusionary rule to searches and seizures performed by the public police. It may be true that the exclusionary rule does little to prevent improper actions not aimed at gathering evidence for use in a criminal prosecution. It may also be true that private police generally care more about protecting their clients than about enforcing the law. But these two functions obviously overlap; this is why there are so many cases like Lima and Taylor, and why courts have had so much difficulty determining whether moonlighting police officers are acting as police officers or private employees. Moreover, the mixed motives of private security personnel do little to distinguish them from the public police, who also spend most of their energies pursuing objectives other than evidence collection. The easy supposition that the exclusionary rule will deter police officers but not security guards is yet another manifestation of the powerful yet erroneous supposition that the public and private realms can be cleanly distinguished based on the functions they perform.

LaFave’s discussion also illustrates how arguments about whether to apply the Fourth Amendment to private security personnel often get mixed up with arguments about whether to apply the exclusionary rule. The confusion is unfortunate because it perpetuates not only the prevailing, largely unexamined notion that the exclusionary rule must logically apply only to constitutional violations, but also the converse assumption that if the exclusionary rule does not apply to private policing, neither does the Fourth Amendment. Considering the two issues separately helps to bring them into sharper focus. But it does not begin to resolve them. For reasons already discussed, the wisdom of admitting evidence seized illegally by private security personnel is far from apparent. Similarly, setting aside the question of remedy does not make any easier the question when, if ever, the Fourth Amendment should apply to private policing. The most obvious
manifestation of state action in a criminal case—the state’s introduction of evidence to secure a criminal conviction—no longer is part of the picture. But the state remains present in other ways deemed significant by the Supreme Court’s state action jurisprudence: it delegates arrest and detention authority, it enters into arrangements for cooperative policing, and it cedes to private firms responsibilities often viewed today as quintessentially governmental. Disaggregating state action doctrine clarifies the choices to be made, but it does not make those choices for us. Nor does it provide a way to avoid the inherent inconsistencies of the public-private divide.

C. Provisional Lessons

1. Doctrinal Disorder and Dialectic Development

In the previous pages, I have tried to show the difficulty of drawing a principled state action line in constitutional criminal procedure, and more particularly of reaching a principled determination regarding the proper treatment of private security personnel. For many this difficulty will be unsurprising—yet another failure of the public-private distinction. Others, however, may have harbored hopes that in criminal procedure, at least, the distinction is unproblematic. Alas no. The Supreme Court’s state action jurisprudence fails to provide firm reasons for distinguishing private police either from public police or from the public at large. State courts and lower federal courts have resolved the problem more out of habit than out of principle. Disaggregating the problem makes many things clearer, but not the answers.

One response to this predicament would be to jettison the state action doctrine altogether in criminal procedure. I think that would be the wrong response, for three reasons.

First, if aimed at ridding criminal procedure of the troublesome distinction between public and private, attacking the state action doctrine would be futile. As Professor Seidman has stressed, the notion of a private realm, beyond the limits of government control and government responsibility, is integral not only to the state action doctrine, but also to the very notion of individual rights. This is manifestly true of the Fourth Amendment, which the Supreme Court has interpreted to protect “reasonable expectations of privacy,” and the Fifth Amendment, which the Court has construed to safeguard an area of autonomous choice. It is true as well of the entrapment doctrine, which blocks prosecution only when “the criminal design originates with the officials of the government.” For better or worse, the public-private distinction permeates the law of criminal procedure. The whole edifice of protections assumes there are areas of private activity worth protecting, in part because the government is not implicated in them. And criminal procedure makes sense only against the background of substantive criminal law, which itself builds on the public-private distinction through pervasive reliance on the concept of individual guilt. That concept is incoherent without “the notion that individuals can make some decisions without those decisions being attributed to the government.”

Second, despite the difficulty of drawing a principled boundary between government conduct and private conduct, placing special restrictions on investigative activity carried
out by the state makes sense. The police do differ from snooping, trespassing, or violent neighbors. They are better organized, they carry more weapons, they can get warrants, and they have closer ties to the government officials on whom we depend to regulate snooping, trespassing, and violence. For similar reasons, despite the recent growth of private policing, public law enforcement still poses risks different from those posed by private security.

Of course, the question still remains precisely where to draw the line between public and private policing. Should “specially commissioned” officers be treated as public police, even when privately employed and supervised? The courts have generally said yes. Should moonlighting officers be treated as public police, even when out of their official uniforms? Again, the courts have generally said yes, although only after engaging in surreal discussions about whether particular officers were “acting as” police officers or as security guards at the times in question. Recognizing that the line between public and private policing is necessarily formal, the ultimate results reached in these cases—minus the surrealism—make a certain amount of sense: individuals with official law enforcement status are distinguished from those without such status. There may be other, more sensible places to draw the line. But as long as we keep in mind the arbitrary nature of the line (which is what the surrealism tends to obscure), the precise location becomes significantly less important.

Indeed—and this is the third and final point—the arbitrary nature of the boundary between public and private policing could itself be a source of doctrinal vitality. We tend to assume that inconsistency is bad, and that the less inconsistency in our law, the better. But this is a prime example of the problem of the second best. A flawless legal system, if there could ever be such a thing, would be wholly consistent. Inconsistency is not just a sign of imperfection; it is intrinsically undesirable for a legal system that views equal treatment as an element of fairness. It does not follow, though, that among imperfect legal systems the most consistent will generally be the best. There are virtues other than consistency. One of those virtues is improvement over time, and the great lesson of the common law is that inconsistency, if used to advantage, can spur doctrinal development. The arbitrary line between contract and tort, for example, allowed courts to experiment with different approaches to similar problems, and to borrow on one side of the doctrinal divide from approaches that had succeeded on the other.

The rules governing public and private police are perfectly suited for this kind of dialectic development. The two groups resemble each other in many ways and present many of the same risks. Moreover, because they increasingly share the same personnel, it is often difficult to tell them apart. Nonetheless the legal regimes under which they operate differ strikingly. And the two legal regimes do not just differ randomly: each contains elements the other is commonly urged to adopt. Critics of the private security industry call for subjecting it to the same restrictions as the public police. Critics of criminal procedure law call for regulating the police under an open-ended standard of reasonableness, interpreted and applied by juries in civil actions for damages—precisely the system we have for private security. The situation is ideal for cross-fertilization.
There has been no cross-fertilization, however, because private security has been largely ignored, and not just by academics. Legislators have shown little interest in reforming the rules governing the private police, or even in collecting basic information about their operations. As a result, we are in a poor position to assess how well the tort system has worked to deter abuses by private security personnel, and no one has tried. We do not know whether tort suits adequately deter the private police, nor whether they overdeter. We need to find out: not just to determine whether and how the system should be reformed—perhaps by offering attorneys’ fees to successful plaintiffs, perhaps by adopting a version of good-faith immunity, perhaps by importing rules of evidentiary exclusion from constitutional criminal procedure—but also to see whether it has anything to teach us about the sensible regulation of the public police.

2. The Illusion of Functionalism

To say that the state action doctrine should be retained is not to say that it cannot be improved. One benefit of examining the doctrine in the context of criminal procedure—the context almost always overlooked by constitutional scholars—is that certain defects become easier to see. Prominent among these is the illusion of a certain kind of functionalism in state action jurisprudence.

The kind of functionalism I have in mind purports to identify state action chiefly by reference to the purposes pursued. Some kinds of purposes are imagined to be inherently public, others inherently private. To distinguish between state action and private conduct, we are told to look not to the formal status of the participants—who employs them, what titles they hold—but rather to what they are doing. This kind of thinking lies behind the public function doctrine, and it motivates the frequent assertion by courts that the legal status of private security guards depends on what roles the guards are performing.

Functionalism of this variety enjoys widespread support among scholars because it promises an escape from a kind of legal formalism widely derided as superficial and obfuscatory. But functionalism in contemporary state action jurisprudence appears itself to be little more than a charade. The public function doctrine has been hemmed in so tightly that almost nothing qualifies as a public function. As we have seen, the doctrine is almost certainly inapplicable to crime fighting and peacekeeping, notwithstanding the common view of the police officer as the paradigmatic state actor. And despite what courts say, the legal status of private security guards rarely turns on the purposes the guards are pursuing; what matters far more is whether the guards have some special, formal status as law enforcement personnel—either because they are moonlighting police officers, or because they have been “commissioned” or “deputized” by some legal authority.

The example of private policing offers some clues why functionalism in state action jurisprudence promises so much more than it delivers. Among the lessons taught by the history of policing is that social understandings often are structured less around rules than around roles, less around categories of actions than around categories of actors. In particular, the notion of the state officer as a distinct and salient legal entity has shown remarkable persistence, notwithstanding large shifts in what policing is thought chiefly to
involve, and a surprising vagueness about what powers officers may exercise. We might say that the concept of the police runs deeper than any particular conception. Any effort to define policing solely in terms of function divorced from form is thus bound to falter, because it fails to account for certain deep aspects of the way we think. This may explain why, when the Supreme Court in *Griffin* found state action in a uniformed security guard’s ejection of black teenagers from a Maryland amusement park, the Court relied on details—the guard’s badge and his status as a “special deputy sheriff”—that have struck some readers as little more than technicalities. Similarly, it may explain why lower courts have refused to resolve state action questions “by looking at the nature of the activities performed by security employees,” and instead have stressed their formal status, explaining that “[a] security guard is not a police officer.” And it may help to explain the persistence of formalism in state action doctrine more generally. Linked as it is to the public-private distinction—and thus to the very notion of individual rights—the concept of the state, like the concept of the police, may run deeper than any particular conception. If so, efforts to define state action solely in functional terms are probably futile.

Whether state action functionalism is worse than futile is a more difficult question. Functionalist rhetoric, even if it seldom becomes much more than rhetoric, could serve as a useful reminder that there are aspects of reality that formalism obscures. On the other hand, by suggesting that state action jurisprudence is less formalistic than it actually is, functionalist rhetoric may lead us to question the contours of that jurisprudence less than we should. Possibly it also reinforces the mistaken and dangerous notion that conduct that does not count as state action typically touches little if at all on matters of public concern. This notion, in turn, may have more than a little to do with the general neglect scholars and legislators have shown to the problems of private policing; it may also explain why the rare attention the legal academy does pay to private policing focuses almost exclusively on the state action problem.

How we assess these countervailing possibilities depends on what kinds of reform we find most promising. The public-private distinction and the notion of individual rights are human constructs, sustained as well as reflected by legal doctrines such as those encountered in state action jurisprudence. Those who hope these constructs can ultimately be discarded may find that the slight oppositional potential of state action functionalism justifies retaining it. But others may conclude, in part based on the history of policing, that the public-private distinction is here to stay, and that we had better learn to live with it. If that means, in part, being as sensible as possible in constructing the formal categories of state and police, and in determining what the legal import of those categories should be, then rhetoric that falsely promises an escape from formalism probably does more harm than good.

**IV. Beyond State Action**

The state action problem holds an enduring fascination. The stakes are weighty, the doctrine transparently makeshift, and the results, at times, strikingly incongruous. In criminal cases, moreover, the question of state action has the added allure of the unexplored, having attracted surprisingly little study. I have tried to show that this neglected aspect of the problem warrants more attention than it has received.
But I have also tried to show that efforts to “solve” the state action problem—to draw a sharp, principled boundary around the activities constitutional law should attribute to the government—are no more likely to prove successful in criminal procedure than elsewhere. Any line will be arbitrary. And there is a danger that questions about the line’s location will distract us from other, more important challenges raised by the spread of private policing.

I end this Article by sketching two such challenges, both for scholars and for the courts. The first concerns the need to know more about the private police, not only to determine what rules should govern them, but also to help assess proposals for reforming public law enforcement. The second concerns the implications of police privatization for the common notion that the chief duty of government is to provide some level of protection against private violence.

A. Private Criminal Procedure

Several decades ago legal scholars began noticing they did not know enough about the everyday world of law enforcement. They knew a lot about indictments, arraignments, and criminal trials, but little about the police—the entry point into the criminal justice system, and for many suspects the most important component of the process. From the 1930s through the 1960s it grew increasingly apparent that serious thinking about criminal justice reform required serious thinking about the police, and that serious thinking about the police required much better information about the everyday world of law enforcement. Governments needed to keep better statistics on police activities, and researchers needed to carry out independent, empirical studies of how the police operated.

To an impressive extent, these goals have now been realized. The federal government collects and disseminates a vast array of police statistics, and many states do so as well. There is a rich literature on the sociology of policing, both empirical and theoretical. Partly as a result, when lawyers and scholars now use the phrase “criminal procedure,” they are generally referring to police methods and the rules that govern those methods, not to courtroom processes. Obviously this usage stems in part from the great growth, particularly in the 1960s, in the amount and complexity of constitutional rules governing the police. But that growth itself was fueled in part by an increased understanding of what the police do.

Much about policing is still shrouded in secrecy or obscurity. But compared to the situation at midcentury, our knowledge today about the police is more striking than our ignorance. And the knowledge we have is important. It has permitted all kinds of discussions—about community policing, about investigative practices, about the Fourth and Fifth Amendment exclusionary rules, and so on—to be carried on with a degree of empirical sophistication that fifty years ago would have been unimaginable.

That kind of sophistication, however, remains impossible with regard to private policing. Indeed, we know less today about private policing than we knew in 1930 about public law enforcement. Our ignorance has made it difficult for us to think seriously about the
rules that govern private policing. Those rules are important because they regulate a world of policing—of private criminal procedure—larger and faster growing than public law enforcement.

Private security, moreover, serves as an entry point not only into the public system of criminal adjudication, but also, and perhaps more frequently, into an alternative system of private adjudication. The sanctions in this private system range from dismissal or ejection, to a return of purloined merchandise, to fines or restitution extorted by the threat of a criminal complaint. The adjudicatory processes preceding these sanctions tend to be informal and inquisitorial; the differences between public and private policing pale beside those between public and private adjudication. And private adjudication is far less visible than private policing. If we know little about the private police, we know even less about private adjudication. One way to start learning is by studying the operations of the least hidden component of the private adjudication system—the private police.

The parallel worlds of private criminal procedure and private adjudication are important not only in their own right, but also because they offer opportunities to assess the wisdom of certain aspects of our public systems of criminal procedure and adjudication. This is particularly true of the remedial rules now in place for the public police. The Fourth and Fifth Amendment exclusionary rules generally do not apply to private guards, nor does the good-faith tort immunity enjoyed by public law enforcement officers, nor the availability of awards of attorneys’ fees to successful tort plaintiffs. All of these rules have been questioned as applied to the public police; in particular, there have been persistent suggestions for replacing evidentiary exclusion with expanded tort liability as the principal means of enforcing the Fourth Amendment. At least two thoughtful scholars have coupled this suggestion with a call to rely more on ad hoc decision making by juries, and less on per se rules promulgated by judges, in assessing the reasonableness of searches and seizures; private criminal procedure complies with this recommendation as well. In several important respects, private criminal procedure thus provides a test case.

The test case is not perfect. To begin with, the variety of differences between the two remedial regimes makes it difficult to isolate the effect of any particular difference. For example, denying fee shifting to successful tort plaintiffs may weaken the effects that would otherwise be caused by denying good-faith immunity to private police personnel. Moreover, as we have seen, the mix of functions performed by private guards today tends to differ from that carried out by the public police, albeit to a degree that remains unclear. But the parallels are close enough to be intriguing. In assessing reform proposals for the remedial regime governing public law enforcement, it would be helpful to know how well the strikingly different remedial regime of private criminal procedure works. This information would be even more helpful if we knew more about how the day-to-day activities of private guards differ from those of the public police.

Empirical studies of the private police, and of the remedial regime under which they operate, are sorely needed. Studies of this kind are difficult to conduct, however, given the industry’s penchant for secrecy. One simple, practical, and important step would be for states to require—as apparently no state currently does—that security firms file
regular, public reports on their activities, with due regard for legitimate privacy interests. Given the growing role of private policing, there is scarce excuse for allowing security firms to continue to keep secret, for example, the number of people they accost and detain, when and how those individuals are questioned, how many suspects are turned over to the police, the number and type of searches conducted by private security personnel, or the circumstances and results of those searches. With or without such data, though, there will be no substitute for close, nuanced field studies of the private police, the sort of thing that we now have in abundance for public law enforcement.

Another aspect of private criminal procedure also warrants study. We need to know more not only about the activities of the private police and about the effectiveness of the rules governing those activities, but also about the manner in which those rules have evolved. For private criminal procedure differs in another salient way from the more familiar world of public criminal procedure: the rules of the former are almost exclusively a matter of state statutory law. Private criminal procedure is neither federalized nor constitutionalized. It therefore offers an opportunity to test the persistent suggestion that the Supreme Court has stunted the progressive improvement of the rules of public policing law by turning the Bill of Rights into a “code of criminal procedure,” and thereby blocking legislative innovation and experimentation.

Innovation and experimentation have not been notable in the regulation of the private security industry, despite complaints that have remained remarkably constant. For the past three decades it has been evident that price competition in the private security industry has kept wages low, turnover high, screening cursory, and training minimal. Media investigations regularly report rampant incompetence and criminality in the ranks of the private police, and every national study of the private security industry has called for stiffer regulation of the screening and training that guards receive. Nonetheless the last such study found that little had changed over the course of two decades. Ten states still leave private security entirely unregulated, only two fewer than in 1971. As of 1990, only four states required as much as twenty-four hours of preassignment training. The typical guard received only four to six hours, and many received none.

The apparent stagnation of regulatory efforts directed at private security may not mean that critics of constitutionalized criminal procedure are wrong. The problems posed by the public police differ from those posed by the private police; if the former are more pressing, they may well be more amenable to legislative solution. But this is not obvious. Once again, discussions about the rules governing public law enforcement—in this case, the wisdom of making those rules a matter of federal constitutional law—might benefit from careful study of aspects of private criminal procedure—in this case, how and why legislation has failed for so long to address such widely perceived problems with the private security industry.

B. Minimally Adequate Policing

The recent, dramatic growth of private policing offers occasion for revisiting not just the relationship between the rules governing public and private policing, and the line the law draws between those two domains, but also, and more fundamentally, the overall mission
of criminal procedure as a field of scholarship and jurisprudence. Since “hiving off” from constitutional law in the middle of the past century, criminal procedure has been overwhelmingly focused on issues of fairness writ small: the procedural obligations owed by the state to an individual suspected of crime. That is to say, it has been all about what content to provide the Due Process Clauses of the Fifth and Fourteenth Amendments in the context of criminal investigation and prosecution. The content provided has been famously elaborate. The Supreme Court has read into the Due Process Clause of the Fourteenth Amendment the specific protections of the Fourth, Fifth, and Sixth Amendments, and has read into those protections an extensive code of police and prosecutorial conduct. As perhaps befitting a law of due process, that code of conduct has been concerned almost exclusively with individualized fairness to suspects and defendants, not with questions of distributional justice.

No comparable body of law has developed regarding the equitable allocation of criminal justice resources. The Supreme Court has refused to recognize a right to minimally adequate protection under the Due Process Clauses, reasoning that the clauses guard only against injuries directly inflicted by government. Due process limits government’s “power to act,” but does not “guarantee . . . minimal levels of safety and security.” The Fourteenth Amendment, of course, promises not only “due process of law” but also “the equal protection of the laws.” In dicta, the Court has read this second promise to prohibit a state’s selective denial of “protective services to certain disfavored minorities.” But the Court has also made clear that the Equal Protection Clause generally prohibits only decisions made with “discriminatory purpose”—that is, “‘because of,’ not merely ‘in spite of,’ . . . adverse effects upon an identifiable group.” As a result, equal protection generally offers no more help than due process in challenging inadequate policing.

State tort law is equally unavailing. The “overwhelming” weight of case law “reject[s] liability based on a general failure to provide police protection.” The doctrinal basis for this result varies. Some courts describe it as a matter of governmental immunity; others declare simply that the state has no duty to furnish police protection, or that the duty is owed to the public at large, and not to any individual, so it cannot provide the basis for a tort claim by a particular plaintiff. The formulation varies, but the result does not: state tort law provides no remedy for inadequate policing.

Underlying this result is a belief that the level of police protection is an allocative question best left to the political branches. To a great extent this notion also underlies the Supreme Court’s refusal to recognize a due process right to safety and security—although the Court justified that refusal chiefly by an appeal to original intent. The idea that adequate policing is best left to politicians rests in turn on three commonly advanced arguments. The first is that courts are poorly equipped to determine the amount of funding the public should spend on any particular service because they lack sufficient information about the budgetary trade-offs. The second is that justiciable standards for the adequacy of public services are hard to find. The third is that judicial involvement is unnecessary because elected officials will have a strong political incentive to provide adequate levels of police protection.
Police privatization seriously undermines this third argument. Even when most policing was public, of course, there was reason to doubt that normal political processes would ensure adequate protection of the poor as well as the rich. But private policing greatly exacerbates the problem. When policing is public, inequalities in the level of protection require conscious, explicit decisions: patrol strength is cut here and increased there. It is thus politically cumbersome for wealthy taxpayers to boost the funding of police protection only for themselves. Private policing makes it simple—indeed, almost inevitable. Why should Bel Air residents vote for higher taxes to fund police cars throughout Los Angeles, when they can—and already do—hire private patrols for their own neighborhood? One might expect a vicious cycle: the wealthy rely on private security, so they are less willing to pay for public police, so public policing deteriorates, so the wealthy further increase their reliance on private security. There is some evidence that the cycle has already begun, at least in Southern California.

This kind of self-reinforcing “secession of the successful” also has affected public services other than policing; one thinks in particular of education. But the problem with regard to policing is especially troubling, and not just for those who view protection against crime as “the most basic function of any government.” When support is slashed for public schools, the results can be readily apparent in quantitative terms—fewer teachers, fewer textbooks, etc. Diminished expenditures for police protection, however, can be reflected not in an apparent decrease in policing, but in a qualitative shift in police strategies from prevention and protection to after-the-fact enforcement. The latter strategy will not make inner-city residents safe and secure, but it may satisfy voters in surrounding communities that enough is being done. If so, “[t]he rich will be increasingly policed preventively by commercial security while the poor will be policed reactively by enforcement-oriented public police,” and “both the market and the government [will] protect the affluent from the poor—the one by barricading and excluding, the other by repressing and imprisoning.”

Since the political process may be unable to avoid this dismal prospect, judges and scholars should begin to give fresh thought to whether law has a role to play in assuring minimally adequate police protection. The difficulties here are daunting. Courts may well have significant institutional disadvantages when addressing budgetary questions, including the problem of developing standards that seem sufficiently judicial. Among the questions that need to be confronted are these: If there is to be a right to adequate policing, is it best located in tort law, statutes, state constitutional law, or federal constitutional law? Should the right give rise to injunctive relief, to money damages for injuries suffered because of inadequate policing, to both, or to neither? And, most fundamentally, what constitutes minimally adequate policing?

This last question will prove particularly difficult. Not only is the question one of degree, but there is considerable controversy regarding whether police resources affect crime rates at all. Summarizing the conventional wisdom among many criminologists, David Bayley calls the notion that current police strategies can decrease crime “a myth.” And if the evidence on the effectiveness of the public police is largely discouraging, there is almost no reliable data on the ability of private patrols to reduce crime. Still, a growing
consensus of criminologists, including Bayley, believe that the police can cut crime if they shift toward some form of “community policing”—a catch-all term that may boil down to “treating a neighborhood the way a security guard treats a client property.” In part this may be because when residents feel safe, they themselves are more likely to act in ways that reduce crime: walking at night, offering assistance to strangers, etc. And, of course, feeling safe has other advantages as well. Plainly, though, adequate policing cannot be merely a question of the number of officers, let alone the amount of money.

Courts and scholars exploring the possible contours of a right to minimally adequate policing might draw lessons from the history of school finance reform litigation, in which similar problems have already been confronted. The effectiveness of school finance lawsuits remains controversial, but the best evidence suggests that at the very least they have succeeded in drawing public and legislative attention to problems of educational funding, attention that in several states has led to meaningful reform.

Duplicating this success in the field of policing will be anything but straightforward, in part because school finance litigation has relied heavily on state constitutional provisions guaranteeing that public education will be “equal,” “uniform,” or “efficient,” or that it will meet some minimal standard of adequacy. Similarly specific provisions about police protection are hard to find. Doctrinal hooks are available, though, if courts and scholars take seriously the often-expressed notion that providing safety and security is “the most basic function of any government.” Prominent among these hooks are the federal and state guarantees that all citizens will receive the equal protection of the laws, guarantees that can be understood—and may originally have been understood—to include some minimally adequate protection against private violence. Today, equal protection is generally understood to require only nondiscriminatory disbursement of benefits or burdens, no matter how minimal the benefits or onerous the burdens. But if the proliferation of private security prods us, as it should, to reconsider what aspects of policing deserve special constitutional attention, that reconsideration must include pondering afresh what it means to promise all citizens, poor or wealthy, the equal protection of the laws.

CONCLUSION

Private policing is not just a problem. It offers lessons about forms of policing the public sector has neglected, aspects of social organization scholars have ignored, and doctrinal possibilities judges have left unexplored. We would be foolish to close our ears.

But we would be at least as foolish to listen uncritically. The untrained, underpaid watchmen of the eighteenth century may have deserved more appreciation than they received at the time, and the Bow Street Runners may rate more retrospective praise than they get today. Nonetheless it is worth remembering the reasons both these systems were replaced by professional, public police forces. Private policing poses real dangers, and not the least is that it will teach us the wrong lessons.
§ 4. Alternative Constitutional Proposals


Privatization is now virtually a national obsession. Hardly any domestic policy issue remains untouched by disputes over the scope of private participation in government, with such disagreements surfacing most recently in proposals for Medicare drug prescription coverage. Privatization is also endemic as a matter of administrative practice. Ours is a system in which private actors are so deeply embedded in governance that “the boundaries between the public and private sectors” have become “pervasive[ly] blur[red].” The lack of a clear divide between public and private is in one sense not surprising, given modern government’s extensive regulation of the “private” sphere. Less acknowledged, however, is the extent of private involvement in the performance of government activities. Private entities provide a vast array of social services for the government; administer core aspects of government programs; and perform tasks that appear quintessentially governmental, such as promulgating standards or regulating third-party activities. While this kind of private involvement is longstanding, more significantly—and here reflecting contemporary political debate—it is expanding. Recent privatization efforts, particularly in health care and welfare programs, public education, and prisons, reveal a trend of greater discretion and broader responsibilities being delegated to private hands.

Despite privatization’s political and practical ubiquity, however, recognition of the extensive intermixing of public and private has failed to permeate thinking about U.S. constitutional law. A foundational premise of our constitutional order is that public and private are distinct spheres, with public agencies and employees being subject to constitutional constraints while private entities and individuals are not. Private involvement in government is addressed primarily through the state action doctrine, which inquires whether, in a particular context, ostensibly private parties should be considered “state (or federal) actors” and thereby subject to the Constitution’s strictures. The terms and rigor of state action analysis have varied over the years, but currently the usual linchpin for finding state action is identifying substantial government involvement in the specific private acts being challenged. The underlying presumption is that cases where private actors wield public power are rare and occur mainly when the government tries to hide behind private surrogates whom it controls. Current doctrine pays little attention to whether the government is, in fact, delegating power to private entities to act on its behalf. To the extent private delegations are considered, it is under the rubric of private delegation doctrine, which assesses whether the Constitution’s separation of powers and due process requirements prohibit the government from delegating certain

* This excerpt also includes part of the Introduction to Metzger’s article.
types of powers to private hands. But constitutional law makes no attempt to link the constitutionality of a private delegation to the risk that it will place government power outside of constitutional controls. Private delegations of government power are upheld without examining whether private actions taken pursuant to such delegations will come within the Constitution’s purview.

Privatization can take a variety of forms. In some instances, privatization represents government withdrawal from a field of activity or from responsibility for providing services, as for example when government disbands a program altogether or sells off state-owned businesses. Privatization of this type raises few constitutional concerns because the Constitution rarely imposes affirmative obligations on government. The focus here, however, is on a different and more common model of privatization: government use of private entities to implement government programs or to provide services to others on the government’s behalf. Rather than constituting government withdrawal, this form of privatization is characterized by a sharing of authority between public and private, with the government giving private entities significant control over and responsibility for government programs. This form of privatization is far more constitutionally troubling. By virtue of their role as stand-ins for the government, private entities often wield powers that I argue should be considered governmental in nature. These activities include not simply physical control over others or control over the content of regulatory standards, but also control over third parties’ access to government resources and benefits.

... [C]onstitutional law’s current approach to privatization is fundamentally inadequate in an era of increasingly privatized government. Much of this inadequacy results from current doctrine’s failure to appreciate how privatization can delegate government power to private hands. Admittedly, distinguishing between instances where private entities are acting on behalf of government and wielding government power, as opposed to instances where the government is simply purchasing private services or sponsoring independent private endeavor, is frequently difficult. But the current emphasis on government involvement in the specific act at issue is a very poor basis on which to rest such a distinction because—as an examination of recent instances of privatization and the Supreme Court’s state action cases demonstrates—private entities can wield these powers alongside minimal government involvement.

Indeed, by assigning determinative weight to the presence vel non of government involvement in specific challenged acts, current state action doctrine arguably gets matters exactly backwards. Such a focus denies constitutional protection against private conduct where the government has given private actors broad discretion over operation of government programs. Under the view that control of access to government benefits and services represents an exercise of government power, however, this is precisely the point at which constitutional protections are most warranted. At the same time, current doctrine applies such protections when they are often least needed—that is, when governments exercise close supervision and thus constitutional norms can be enforced by targeting government action directly. Worse still, focusing on government involvement creates perverse incentives for governments to forego close oversight of their private partners,
even though such oversight is an important means for ensuring that private actors adhere to constitutional requirements.

The legal academy increasingly has taken note of the phenomenon of expanding privatization. But for the most part, that scholarship has not addressed the lack of fit between existing constitutional law and the reality of modern governance, other than to conclude that privatization often removes government programs and activities from constitutional scrutiny, or to assess whether particular instances of privatization are constitutional. Little effort is made to rethink the basic terms of constitutional analysis in the face of the disconnect between administrative reality and constitutional doctrine. Instead, the focus is on reforming nonconstitutional law to better address accountability concerns raised by privatization, specifically the moral hazard problem: the danger that private actors will exploit their position in government programs to advance their own financial or partisan interests at the expense of program participants and the public. In response, scholars have proposed reforming administrative statutes to improve public oversight of privatized programs, imposing greater regulation and contractual controls on recipients of government funds, or ensuring program participants’ access to private law remedies.

Developing such measures is surely a crucial part of coming to terms with the new reality of privatized government. Yet reexamining constitutional law’s current approach to privatization is at least as essential. A central premise of U.S. constitutional law is that the Constitution imposes limits on the actions that governments can take. Critically, these limits apply to all exercises of government power, whether wielded by officially public or nominally private entities; in addition, individuals injured by exercises of government power can enforce these constitutional limits in court. This central premise is what I refer to in this Article as the principle of constitutional accountability. As used here, constitutional accountability is not the commonly-invoked idea of political or electoral accountability, nor the idea of programmatic accountability alluded to above, but rather the concept of legal accountability that is basic to our constitutional system. The inadequacies of current state action doctrine mean that private exercises of government power are largely immune from constitutional scrutiny, and therefore expanding privatization poses a serious threat to the principle of constitutionally accountable government. Where governments have adopted legislative or regulatory measures that offer substantial protection against private abuse of power, this constitutional accountability concern appears less pressing. But governments may be unwilling to incur the costs associated with such measures, particularly when the programs being privatized serve politically impotent groups. Addressing the disconnect between constitutional law and modern administrative reality is essential to ensure that basic constitutional protections exist across the board.

. . . The central challenge posed by privatization is not just how to enforce the core principle of constitutional accountability, but how to do so without transferring the political branches’ regulatory authority to the courts. While enforcing constitutional requirements frequently restricts the government’s regulatory role, such a restriction is of particular concern in privatization contexts. The premise of the public-private divide in constitutional law is that the rules governing private actors should be politically rather
than constitutionally determined. Moreover, preserving regulatory flexibility is especially important if governments are to successfully reap privatization’s potential for greater efficiency and innovation while still guarding against private misuse of public power.

The answer to the constitutional challenge of privatization, I argue, is to take a middle way: to insist that private exercises of government power must comport with constitutional requirements, yet recognize that in the context of private delegations, the method of enforcing these requirements, as well as their substance, may differ from when government officials act alone. Under the standard model of constitutional adjudication, a finding of state action means that relief for constitutional violations often runs directly against the government’s private partners. But direct application of constitutional limits to private actors is not necessary to achieve constitutional accountability. Such accountability requires that individuals be able to enforce constitutional requirements against specific private exercises of government power; it further demands that they ultimately be able to assert their challenges in court. Nothing in our constitutional structure or history, however, mandates that this enforcement take the form of judicial imposition of constitutional requirements directly against private actors, as opposed to judicial enforcement of these requirements through other means.

Under the approach advocated here, the crucial constitutional question is whether adequate accountability mechanisms exist by which to ensure that private exercises of government power comport with constitutional requirements. If such mechanisms are lacking, the appropriate judicial response is not subjecting private entities to direct constitutional scrutiny, but instead requiring that the government create such mechanisms as the constitutionally-imposed price of delegating government power to private hands. A central advantage of this approach is that it gives governments an incentive to adopt measures that protect against potential private abuses. Absent mechanisms that adequately substitute for direct constitutional review, delegation of government power to private entities would now be unconstitutional. Equally important, avoiding direct constitutional scrutiny of private actors allows the government greater flexibility because the government can choose among a variety of accountability mechanisms in structuring instances of privatization to meet constitutional demands. . . .

Modern privatized government does not fit easily within the paradigms of U.S. constitutional law. A fundamental tenet of constitutional law posits an “essential dichotomy” between public and private, with only public or government actors being subject to constitutional restraints. With rare exception, the Constitution “erects no shield against merely private conduct, however discriminatory or wrongful.” The reigning constitutional paradigm thus strictly compartmentalizes society into public and private spheres, and does not acknowledge any substantial blurring between the two.

As a result, the move to greater government privatization poses a serious threat to the principle of constitutional accountability. Although not often articulated, this principle also lies at the bedrock of U.S. constitutional law. To begin with, it embodies the core idea of constitutional supremacy and constitutional government, namely that the Constitution imposes restrictions on government that the political branches lack ability to alter. Crucially, these restrictions apply not only when the formal organs of government
act, but also whenever government power is exercised. This broad scope of application reflects the proposition that the Constitution encompasses all “actions of the political body denominated a State, by whatever instruments or in whatever modes that action may be taken,” coupled with the recognition that “the abstract thing denominated a State” can exert its powers only through the actions of persons. Restricting the Constitution’s ambit to apply only when the government formally acts would avoid the difficulties in determining whether a private entity is wielding government power; doing so might also yield programmatic benefits, in that private entities and government could pursue the most efficient and effective forms of program operation unconcerned with constitutional requirements. But such an approach would significantly eviscerate the concept of a constitutionally constrained government. Adequately guarding against abuse of public power requires application of constitutional protections to every exercise of state authority, regardless of the formal public or private status of the actor involved: “It surely cannot be that government, state or federal, is able to evade the most solemn obligations imposed in the Constitution by simply resorting to the corporate form” and thereby transferring operation of government programs to private hands.

Constitutional accountability’s second dimension centers on how these restrictions are enforced. Accountability is often defined in political terms, as the accountability of officials to the electorate. Political accountability is certainly one way of enforcing constitutional limitations, with those who ignore fundamental constitutional constraints being ejected from office. But there is also legal accountability, the focus here, which is concerned with individual enforcement of constitutional restrictions in court through judicial review. A defining characteristic of the U.S. constitutional order is the authority it gives to judges to enforce constitutional constraints against other government officials at the instance of private individuals claiming injury from unconstitutional action. . . .

To say that individual enforceability is basic to our constitutional system, however, is not to say that full remediation is. Immunity doctrines and the Court’s growing reluctance to imply *Bivens* actions make obtaining damages for constitutional violations increasingly difficult. These barriers to damages may lead to underenforcement of constitutional norms, but the principle of effective constitutional accountability is preserved by the availability of injunctive and declaratory relief.

The danger is that handing over government programs to private entities will operate to place these programs outside the ambit of constitutional constraints, given the Constitution’s inapplicability to “private” actors. The four examples of contemporary privatization help explain what this could mean in practice. The administrative agency that handles Medicare claims is subject to constitutional due process requirements; accordingly, it cannot deny requested services without providing some minimum opportunity to be heard. Theoretically, a private MCO is free from such constraints. Similarly, government welfare offices are strictly limited in their ability to consider a welfare recipient’s race or religion, and cannot drug-test program participants without basis to suspect drug use. However, private welfare contractors theoretically are not so circumscribed by the Constitution. While public schools must respect the First Amendment rights of teachers and students, private schools theoretically can fire employees and expel students who question how the school is run. Public prisons are
bound by the Eighth Amendment to provide adequate medical care to those in their
custody, but theoretically private prisons are free to refuse to provide medical care at all.

“Theoretically” is an important qualification here. On rare occasions, courts may hold
that such nominally private action in fact constitutes state action for constitutional
purposes. More frequently, the actions described may run afoul of legislative, regulatory,
or contractual requirements, and the government may itself police the conduct of its
private partners to ensure they adhere to constitutional prohibitions. Tort law also may
provide a basis for recourse against some private actions. In light of such alternative
protections, some might contend that constitutional accountability fears regarding
privatization are misplaced. On this view, preserving a private actor’s nongovernmental
status arguably better ensures accountability because it offers more opportunities for
individuals to recover money damages, from which public entities and employees are
frequently immune. Yet caution is needed so as not to exaggerate the extent to which
private law offers equivalent or greater protection of individuals’ rights than is available
through constitutional means. Statutory and regulatory measures or contractual provisions
may offer more extensive protections, but these protections exist as a matter of legislative
or executive grace, and thus largely can be rescinded or limited where governments see
fit. As a result, these forms of protection standing alone do not substitute for
constitutional constraints, in the way that constitutional constraints function in U.S.
constitutional law, as requirements beyond the power of government to alter and which
individuals have the power to enforce.

A second objection to viewing the absence of constitutional constraints on privatization
as problematic emphasizes the voluntary status of many privatized government programs.
Generally speaking, the Constitution does not impose affirmative duties on government.
Or to put the point more starkly: Those “who wrote the Bill of Rights were not concerned
that government might do too little for the people but that it might do too much to them,”
and the purpose of the Fourteenth Amendment was similarly to “protect Americans from
oppression by . . . government, not to secure them basic governmental services.” Why
should it matter whether government programs become exempt from constitutional
constraints as a by-product of the government transferring these programs to private
hands? Is this situation so constitutionally different from one in which the government
remained uninvolved from the outset, presuming that private charity would meet social
needs? Arguably, providing a blanket constitutional exemption for inaction while
simultaneously holding that privatization raises constitutional concerns simply penalizes
the government for affirmatively addressing social problems. This is not an attractive
postulate for a sensible system of constitutional law.

This objection has force, especially when joined with the idea that government should
have a freer hand when spending public resources than when regulating. Notably,
however, the lack of an affirmative government obligation to act has not meant that when
the government does act, it can do so outside of the Constitution’s purview. This
distinction merits emphasis, for as discussed above, most instances of privatization are
poorly described as simply a return to government inaction. Far from it. Instead, they
usually represent a change in the form of government action, with discretionary authority
over government resources and programs being transferred to private hands. That private
entities play a central role in a program’s implementation may well counsel for application of different constitutional requirements, but the danger is that privatization will remove what are essentially government programs from all judicial constitutional review. The question thus becomes whether it is possible to preserve both the principle of constitutionally constrained, accountable government and constitutional law’s public-private divide in the face of the move to privatized governance.

Yet the threat to constitutional accountability represents only part of the constitutional equation. Applying constitutional norms to the government’s private partners could solve this problem, but at a significant cost. One effect would be to intrude on private autonomy, as constitutional constraints limit the ability of private entities to operate as they see fit. These autonomy concerns might seem less pressing in privatization contexts, given that private entities are reaching out to interact with the state and take on roles in government programs. But the concern is not with preserving a particular entity’s freedom so much as ensuring the independence and vitality of a private civic sphere writ large. Such a sphere not only provides a needed space for individuals to define their own identities, but also fosters development of institutions that offer a potent check against overreaching government. At a minimum, autonomy concerns counsel for caution so as not to apply constitutional constraints where constitutional accountability concerns are otherwise addressed.

Of course, governments could—and do—impose significant restrictions on private entities, particularly as a condition for receipt of government funds. Thus, perhaps more telling than individual autonomy concerns is the effect that extending constitutional requirements to private actors has on the government’s regulatory prerogatives. “Constitutionalizing” the government’s private partners effectively transfers the power to decide what rules should bind these private actors from the political branches of government to the federal courts. This separation of powers concern with expansion of the judiciary’s regulatory role is a central justification for retaining the public-private divide in constitutional law, and it applies no less strongly in the government privatization contexts.

Indeed, the four examples discussed above indicate that preserving the political branches’ regulatory flexibility in privatization contexts is particularly important. Privatization holds the potential to yield more efficient and innovative government programs, by allowing the government to harness private expertise, flexibility, and market competition to its advantage. Yet privatization can also lead to abuse and exploitation, because the financial incentives of private companies and organizations often run counter to the public interest and the interests of program participants. Difficult policy choices are involved in deciding whether to impose particular requirements on how private providers and regulators operate; for example, procedural protections for program participants may guard against private exploitation, but also may translate into higher costs and lessened flexibility, thereby undermining the programmatic benefits obtained from privatization. The government’s ability to tailor regulatory structures to address identified abuses while exploiting private strengths is likely to be pivotal to successful reliance on privatization. Equally important will be allowing governments room to experiment with different
approaches to privatization, so that practical experience can inform regulatory choices and governments can address new problems as they emerge.

Federalism concerns are also at stake here. As the examples above suggest, much of the turn to privatization is happening at the state and local level or in programs (such as Medicaid or TANF) that are joint federal-state endeavors. Extending constitutional requirements to the government’s private partners thus portends significant federal court intrusion into the administrative decisions of state and local governments. Again, the federal government’s ability to impose requirements on state governments, particularly as a condition of federal funding, somewhat limits the force of this federalism objection. But federalism concerns may counsel against judicial imposition of constitutional commands on private actors notwithstanding congressional and executive powers to regulate state and local governments.

Some might argue that adherence to a constitutional system requires that primacy be given to preserving constitutional limits on government power, regardless of the policy tradeoffs involved. This argument is plainly too facile, however, as the individual autonomy, separation of powers, and federalism concerns articulated above also stand as constitutional limits on government power. The argument also errs to the extent that it ignores the way that privatization can enhance constitutional accountability by improving government functioning. If private prisons lead to improved prison conditions and services, that is hardly irrelevant from an Eighth Amendment standpoint. Private professional norms and independent judgment can offer individuals important assurances of fair treatment that meet due process requirements. More simply put, constitutionally-embedded values lie on both sides of the balance. The constitutional challenge posed by privatization is devising a means to preserve constitutional accountability without sacrificing government regulatory flexibility and its associated benefits.


The current challenge is how to update the state action doctrine in a way that preserves the distinction between state and private actions and is still capable of recognizing new forms of activity in the public sphere. This is not an argument against the basic premise of the American Constitution that the government carries a special burden of Constitutional duties, which is not applicable to private actors. It is an argument against the premise that the definitions of government action can be detached from the changes in real life.

Having mentioned Brown v. Board of Education (1953) [as a case where the Supreme Court explicitly recognized that education is “perhaps the most important function of state and local government”], I will try to explain the constitutional challenge posed by privatization in regard to alternative methods to fulfill the responsibility of the state for the education of children. Usually, the states guarantee basic education by the operation of public schools. In these circumstances, the schools are owned by public authorities
(the “state”), and there is no doubt as to their constitutional accountability. Let us assume that as an alternative method of operation (for economic reasons), the state chooses to close its public schools and pay for the services of privately-owned institutions (by financing the enrollment of students). Now, the question is whether there should be a difference in the scope of the constitutional rights of the students in these two modes of operation. Even if private schools participating in privatization projects are not likely to adopt racist policies, there are still other constitutional concerns, i.e., what will be the scope of First Amendment rights of the students (regarding both freedom of speech and freedom of religion); will they have due process rights? Obviously, the total disappearance of the public schools system is not likely to happen. However, similar changes do and will happen, even in the field of education. Although education for the average child is still administered by the public schools system, significant numbers of troubled children are now educated in private institutions, financed and supervised by the state (as in *Rendell-Baker*). These children should not have lesser constitutional rights while in school, simply because the government chooses not to form a special school for them. In *Rendell-Baker*, the Court could avoid the question of the rights of the enrolled students, because the petitioners were employees of the institution. However, it did not form a theory of state action that is capable of distinguishing between the two. In the same year, in *Milonas v. Williams* (10th Cir. 1982), . . . the Tenth Circuit recognized the existence of state action, in a case similar to *Rendell-Baker*, in all respects, except that the plaintiffs were students claiming infringements of their constitutional rights (referring to use of a polygraph machine, monitoring and censoring mail, use of isolation rooms and use of excessive physical force). The circuit court explained that the *Rendell-Baker* decision was not ruling, because of the difference in the identity of the plaintiffs. For reasons already given, the recognition of a state action in the above circumstances was justified. However, it is doubtful, whether it is founded on the precedents of the Supreme Court.

A doctrine of state action for the present and future times must abandon some of the limitations with which it has been burdened. The current doctrine is capable of serving as a starting point. The two theories discussed before: the public function theory and the nexus theory express sound policy considerations, and therefore deserve further evaluation.

Public function theory bases the constitutional accountability of the private entity on the operation of a function considered to be part of the state’s responsibilities. Earlier in the discussion, I argued that the application of the theory should reflect current understandings of the responsibilities of the state, including health care, education and welfare. In other words, the operation in the public domain, as it is perceived today, should serve as the basic layer for the application of the state action doctrine to so-called private bodies.

The public nature of the activity cannot serve as a sole factor, however. If the nature of activity could suffice, then all voluntary organizations operating education and community services would be embraced by the new state action doctrine, for no good reason. More specifically, the private school is not a state actor, despite its educational task, unless it serves as a substitute for a public school, that is when the state does not
offer a public alternative to its services (in this example, a local public school). In other words, to establish a state action, the public nature of the activity is not enough. In addition, it is important to establish a nexus between the private actor and the government. In this respect, the close nexus idea is also not superfluous, although its application needs to be transformed. When a private agency operates a public service for the state, the nexus requirement should be considered as fulfilled. In view of the modern state, it is wrong to base recognition of symbiotic relationships between the state and private entities on the existence of pecuniary profit to the government, as was inferred by the Burton decision. In the usual case, the benefit to the government from cooperation with the private entity is that it is freed from building an administrative apparatus that will operate the same public function. For example, if education is perceived as a public function, and if the state frees itself from opening schools for problematic children by financing their enrollment in private institutions, then these institutions are in fact a substitute for the state. The same goes for private nursing homes accommodating Medicaid residents, for private correctional facilities, or for private welfare agencies operating in similar circumstances. On the other hand, the religious private schools operating throughout the country are not state actors. These schools operate a service which is public in nature—education. But, they are not substitutes for public schools in the sense explained before, since they are only an option to existing public schools.

A closer look at the basic precedent of Marsh v. Alabama (1946) reveals that even there the public function rationale was tied to a supplementary consideration—the private actor serving as a substitute to government activity. The gist of the decision was not only that the management of a town is a public function. It was also that the private corporation managing the town was operating as a substitute for a local municipality. In other words, the state was freed from the burden of supplying municipal services for the people living in the company-owned town. Therefore, the equivalent of the company-owned town is not a private school operating as an alternative to an available public school, but rather a private school operating as a substitute to a public school, when the latter is not available.

In some cases, the private enterprise which serves as a substitute for an official action may enjoy monopolistic status, but not necessarily. This may vary in accordance with the public function that is at stake. Management of a town or an operation of a public utility are monopolistic in nature. Therefore, their so-called private equivalents are deemed to be monopolistic, too. But, this is hardly the case with social services like schools or nursing homes. Therefore, the monopolistic status of the private entity cannot be a decisive test. When exclusivity exists, it may be easier to establish the conclusion that the private corporation operates as a substitute for the government, but not more than that.
How should the concept of state action be understood? What would a coherent jurisprudence look like? The Warren Court seemed to be groping toward a holistic approach that was promising, however, the Burger and Rehnquist courts have retreated from that analytical approach, restricted the scope of the doctrine and arguably further confused the issues.

Any workable state action doctrine will require flexible application, especially during a period in which we are “reinventing” government. Flexibility need not trump consistency and predictability, however. Certain characteristics of the relationship between government and private entities will always be relevant to the inquiry whether an action can be fairly attributed to the state. Among the dispositive elements will be the existence, nature and extent of government funding; the nature and extent of government control of the activity in question; the extent to which government has authorized a contractor to exercise government powers; and a functional (holistic) analysis. All of these inquiries, I submit, should be approached through the lens of conventional agency law analysis.

When the state authorizes a private entity to act on its behalf, it creates an agency relationship. When the agent is authorized to exercise powers that are essentially, even if not exclusively, governmental, we are justified in finding that government has acted through that agent. Well-settled principles of agency and partnership could be particularly helpful and relevant to this analysis, and it is curious that those principles have not been applied, even by analogy.

Agency principles would suggest that when government cloaks a contractor with real or apparent authority to act on its behalf, the ensuing actions should be deemed governmental. Section Seven of the Restatement (Second) of Agency defines authority as “the power of the agent to affect the legal relations of the principal by acts done in accordance with the principal’s manifestations of consent to him.” Section Twenty-Six provides that “authority to do an act can be created by written or spoken words or other conduct of the principal which, reasonably interpreted, causes the agent to believe that the principal desires him so to act on the principal’s account.” Section Twenty-Seven defines “apparent authority” as “conduct of the principal which, reasonably interpreted, causes the third person to believe that the principal consents to have the act done on his behalf by the person purporting to act for him.” While a determination that an agency relationship exists remains heavily fact-sensitive, these are far more reasonable standards than the requirement of “coercive power” or “significant encouragement” demanded by Rehnquist in *Blum*. The issue should be *control*, not coercion.

In an agency relationship, the agent has a duty to follow the principal’s direction even when the principal has previously agreed not to intervene in the agent’s exercise of discretion. Significant legal consequences follow from the conclusion that a particular relationship is one of agency. An agent has power to commit the
principal to transactions within the scope of the actual or apparent authority created by the agency relationship, and the principal is vicariously liable for instances of misconduct committed by the agent when acting within the scope of the relationship.

The law of agency is concerned precisely with the question on which a finding of state action will turn; that is, when is it fair to hold one person or entity responsible for the actions of another?

How would such an inquiry proceed in the privatization context? Where government money passes to a presumptively private entity, a threshold inquiry should be whether the transaction is a purchase of goods or services, which should not constitute state action, or the funding of authorized activities, which may. Purchase involves a product or service that is generally available and relatively standardized, where production of the good or performance of the service is substantially, if not entirely, controlled by the vendor. When the state goes into the market looking for computer support or engineering services, for accountants to perform an audit, for asphalt to use in paving projects, or similar widely traded goods and services, it is relatively clear that government is simply making a purchase. Even where the transaction is apparently a purchase of services, however, a significant long-term relationship between a contractor and the government, where the government’s business constitutes a majority of the contractor’s income, should be held to raise a rebuttable presumption of agency, or state action. When a single customer accounts for most of a contractor’s income, that customer clearly has the power to control behavior. The existence of that leverage justifies raising the presumption, which can be rebutted by evidence that no active participation by government in fact occurred.

Where government controls the manner in which work is done, there should be a finding of state action, without requiring a finding of explicit authorization of the disputed act. This rule is consistent with the Court’s current articulation of the law, if not its practice. Mere regulation should continue to be insufficient, but something less than direct coercion should implicate the State. Where a regulatory scheme substantially limits the options available to the private entity, or prescribes goals and delimits acceptable methods for achieving them, the state should bear responsibility for the results. In addition to money or payment, indicators relevant to the issue of control include ownership, relationship, degree of regulatory supervision and authority, and the actual course of dealing between the parties. Courts might helpfully analogize with existing tests to determine whether a worker is an employee or an independent contractor. While independent contractors can also be agents, certain of those questions can provide a useful analytical threshold.

Each of these approaches attempts to answer the question whether the private entity is acting as a proxy for government under the facts of the case. If the contractual relationship has replaced government employees who were previously providing the service, common sense suggests the contractor is a government proxy. Where government is responsible for delivery of the service, there should be at least a rebuttable presumption of state action. The “exclusive function” test is not good public policy,
because the issue is not whether the activity is one that only government does or has ever done, or should do. The issue is whether it is government that is actually “doing” the activity in question. Where government undertakes an activity, funds it, authorizes a contractor to act on its behalf, and effectively dictates the manner in which it is done, that activity should fairly be attributed to government.

Courts have been reluctant to burden governmental units—and ultimately taxpayers—with liability for the actions of private contractors, but the concern is arguably misplaced. Government can protect tax dollars by contracting for hold harmless or other indemnification provisions. Liability is a recognized cost of doing business, and the allocation of costs is a proper subject for contractual negotiation. Indeed, a refusal to allow government to evade its constitutional responsibilities through contracting will force an explicit recognition and accommodation of potential liability costs. Such a result would benefit everyone: private contractors, government units, and most of all citizens who have a right to demand fiscal, political and legal accountability from those they elect to office.

When the government acts, it should be accountable. The instrument government chooses should not alter that result. Whether government delivers drug counseling, job placement, or any other service through a state agency or a faith-based organization, the program is state action. Courts should recognize that reality and require adherence to constitutional standards.

§ 4.4. Notes and questions

1. **The nuanced approach: cumulativeness.** As Metzger (as well as others) points out, Sullivan is more formalist, while Brentwood Academy is more nuanced. Some argue that the solution is to follow a more Brentwood Academy-type approach and consider all factors cumulatively, rather than treating them one by one. See Krotoszynski, Back to the Briarpatch: An Argument in Favor of Constitutional Meta-Analysis in State Action Determinations, 94 Mich. L. Rev. 302 (1995) (writing of course before those two cases were decided). Would this solve the problem? Or would a cumulative approach still allow judges who were disinclined to find state action to follow their disinclinations? Moreover, wouldn’t this worsen the problem of overinclusiveness (see note 5 of the Notes on the “quasi-governmental organization” cases)? Metzger, at 1451–52.

2. **The positivist approach: social understandings.** Barak-Erez proposes to beef up the “public function” theory by replacing its emphasis on exclusive state responsibilities with “current understandings of the responsibilities of the state, including health care, education and welfare.” By “responsibilit[y] of the state” (which she also refers to as “the operation in the public domain”), Barak-Erez apparently meant not anything the state happens to be doing now (which would include running zoos and cafeterias), but only (perhaps) what the public recognizes as an entitlement or “core” government service. How can we tell what
the public supports as an entitlement, and what it merely supports as convenient policy? (For instance, the public may support welfare for the “truly needy” but not for the “able-bodied”; actual welfare policy may draw the line somewhere in between. Is all of welfare provision then a public function, or only the subset of it that is recognized as an entitlement?)

More fundamentally, do such shared understandings exist, or is this wishful thinking? Note that the 2008 Republican Party platform “support[s] the private practice of medicine and oppose[s] socialized medicine in the form of a government-run universal health care system” (and there have been recent proposals for cuts in medical coverage); the school choice movement can be read as supporting government funding of education as an alternative to government provision of education; and the 1996 welfare reform law denies any individual entitlement to TANF welfare services (42 U.S.C. § 601(b)). (Note the analogy to other (not uncontroversial) areas of the law: the Katz v. United States (1967) test for Fourth Amendment searches looks to “reasonable expectations of privacy” by consulting social understandings, and the Kaiser Aetna v. United States (1979) test for Fifth Amendment takes looks to “reasonable investment-backed expectations.”)

3. The positivist approach: statutory responsibilities. How about limiting public functions to functions that the government has committed itself by statute to provide? Cf. Marshall’s dissent in Rendell-Baker, where he stresses that special-needs education is a statutory responsibility of the state of Massachusetts; see also note 3 of the Notes on the “quasi-governmental organization” cases, supra.

But since anything the government does is presumably authorized by statute, would this include, as suggested in note 2 supra, everything the government currently does, including running cafeterias, in which case private restaurants would be covered as well? If the government committed itself to providing relief for the poor, would private poor-relief agencies (even ones not funded by the government) be performing public functions? What if the government service is limited in scope or time—do private providers of the service cease to perform a public function once they exceed that scope or time?

Note that if this is the case, there would be no special, barely-regulated status for government subsidies of private providers, since these private providers (to the extent they were doing what the government had committed itself to do) would already be constitutionally regulated. See Metzger, at 1450.

What if public functions were limited to functions that the government has committed itself to constitutionally? (This would generally be a small set.) What about functions that the government is constitutionally required to provide (even if not explicitly, and referring both to the U.S. Constitution and, if applicable, to the state constitution)? This may explain the result in West (again, see note 3 of the Notes on the “quasi-governmental organization” cases, supra), but again may be underinclusive, in light of DeShaney.

242
Draft — do not distribute!
4. **The non-positivist approach: an objective search for core government functions.** Perhaps one could ignore both popular views and statutory enactments, and identify governmental functions philosophically? The Supreme Court has abandoned the effort to identify core government functions as “unsound in principle and unworkable in practice.” *Garcia v. San Antonio Metro. Transit Auth.* (1985). One could of course use one’s favored philosophical approach to determine what is governmental, and this will work, provided the philosophical approach is determinate enough (for instance, one could limit “governmental power” to instances of “coercion” according to one’s favorite theory); but note that judges are unwilling to explicitly admit their allegiance to any comprehensive system of political philosophy, and such an admission would be in tension with prevailing views of the judicial role. Cf. *Lochner v. New York* (1905) (Holmes, J., dissenting). Nonetheless, you may conclude that referring to some philosophical approach outside of the Constitution is unavoidable.

5. **Public functions as decisions affecting eligibility for government benefits.** Metzger grants that “[i]dentifying what constitutes government power is a notoriously hazardous enterprise, and little agreement exists on where the boundaries of government power as opposed to private power lie.” Metzger, at 1396–97. However, she suggests that “control over government resources is commonly thought of as government authority,” and that such control includes “denials of medical services by Medicare and Medicaid MCOs,” welfare contractors’ assessment of “a welfare participant as not needing additional job training,” or voucher schools’ suspicion of students for disciplinary infractions. Id. at 1399.

But “Justice Brennan’s *Blum* dissent suggests caution, however. He argued that doctors treating patients under Medicaid should not be considered state actors simply because their treatment decisions determine how much and what kinds of benefits Medicaid participants receive. In other words, almost any decision by a private entity—whether to provide certain medical treatments, whether to recommend a job training program or instead require a beneficiary to work off her welfare benefits in community service, whether to require a child to repeat a grade—in fact qualifies as determining eligibility for government benefits when private entities are providing services on the government’s behalf.” Metzger, at 1450–51.

Brennan’s *Blum* dissent suggests that there are different ways to “determin[e] eligibility for government benefits.” If the government staffs its welfare program with private contractors, this clearly seems to be government power. Now consider the following example: You are issued a food stamp that says “Not to be used for junk food.” Retailers accept food stamps because they expect government reimbursement when they send them in later. If you try to use your food stamps to buy Twinkies, the grocer refuses to accept the food stamps and requires cash. If you try to use them to buy a chicken, the grocer accepts them. If you try to use them to buy Wheat Thins, the grocer accepts them, or not, based on his assessment of whether the government will consider Wheat Thins to be “junk...
food.” Is the grocer exercising delegated government power when he makes this decision? That is, does he owe you constitutional due process if he refuses to accept your food stamps for Wheat Thins? Compare this hypothetical with the facts of *Blum*.


While few would dispute the need to ensure that constitutional doctrine reflects the reality of privatized government, the question remains whether it is possible to achieve that goal without simultaneously destroying regulatory flexibility. The solution to this quandary, I suggest, is to return to the road not taken and to rethink state action in private delegation terms. Under a private delegation approach, the key issue becomes not whether private entities wield government power, but rather whether grants of government power to private entities are adequately structured to preserve constitutional accountability. Provided that alternative mechanisms exist to ensure that government power ultimately stays within constitutional limits, exercises of government power by constitutionally immune private actors do not present constitutional concerns. This approach secures constitutional accountability by ensuring that individuals are able to enforce constitutional limits on government power; but it also grants government more flexibility by allowing choices of how best to preserve constitutional limits to be made by the political branches in the first instance.

This Part is devoted to justifying and constructing such a private delegation analysis. I argue that the central criterion for singling out particular private delegations for enhanced scrutiny is whether they authorize private entities to act on the government’s behalf, a factor usually established by assessing whether the requirements of agency are met. Even where agency relationships exist, however, mechanisms such as close government supervision or the presence of a meaningful choice among providers are often sufficient to preserve constitutional limits on government power. After discussing how unconstitutional delegations should be remedied, I then assess the extent to which this proposed private delegation approach is compatible with existing constitutional law and the likelihood of judicial adoption. Finally, in order to give a clearer illustration of what it would mean in practice, I apply the proposed analysis to . . . four recent examples of privatization . . . .

A. The Case for Rethinking State Action in Private Delegation Terms

Privatization appears to force a stark, binary choice between constitutional accountability, on the one hand, and political control and regulatory flexibility, on the other. Expanding state action analysis leads to judicial imposition of constitutionally-derived codes of behavior; constitutional constraints on government power are thereby preserved, but at
the cost of significant intrusions on the political branches’ regulatory role and on the opportunity for regulatory experimentation. But is this conflict in fact so unavoidable?

Obviously, constitutional requirements operate to restrict the government’s freedom of action, yet the relationship between these two need not be a zero-sum game, as it is under current state action doctrine. To avoid this result, however, it is necessary to cast aside state action’s understanding of constitutional accountability as requiring that constitutional constraints apply directly to every exercise of government power. This understanding is excessively narrow. . . . [C]onstitutional accountability demands only that individuals be able to ensure that exercises of government power stay within constitutional limits. Direct application of constitutional constraints to private actors wielding government power is certainly the most common means of preserving such limits. But no reason exists to conclude that this means is constitutionally mandated; other nonconstitutional mechanisms that impose adequate limits on exercises of government power can also suffice.

The Court’s private delegation decisions emphasizing the role of government oversight provide a valuable lesson. True, the Court did not link such oversight to the question of whether private delegates’ actions should be subject to constitutional scrutiny, or probe beyond the formal structures applicable to a delegation to determine whether the private delegate was in fact meaningfully constrained. Nonetheless, these decisions contain the important insight that the structure of a private delegation should matter more in determining its constitutionality than the mere fact that private actors are exercising government power. More specifically, they recognize that nonconstitutional accountability mechanisms can adequately address the constitutional concerns that private delegations might otherwise create.

This idea that alternative accountability mechanisms can satisfy constitutional requirements is one that surfaces elsewhere in constitutional law. For example, the availability of ex post tort actions against the government to recover for injuries to property or liberty may satisfy procedural due process requirements, obviating the need for administrative hearings. Similarly, in *Miranda v. Arizona* (1966) the Court laid out specific warnings that must be given prior to custodial interrogations, but indicated that these warnings were not required if “other fully effective means are devised” to protect suspects’ Fifth Amendment rights against self-incrimination. More recently, . . . *Zelman v. Simmons-Harris* (2002) . . . underscored the importance of “genuine and independent choices of private individuals” in concluding that the use of public funds to pay tuition at religious schools did not violate the Establishment Clause.

A focus on alternative accountability mechanisms also accords with modern separation of powers jurisprudence, where the Court generally eschews insistence on formal restrictions for a more flexible inquiry into whether sufficient checks exist in practice to prevent excessive accumulation of power in one branch. Indeed, the separation of powers parallel runs even deeper. The Court’s functionalist bent in the separation of powers arena reflects in part its recognition of modern administrative realities; regulatory complexity and uncertainty (not to mention the institutional constraints of Congress) mean that much of the content of governing rules must inevitably be set by administrative
agencies. In like vein, focusing on alternative mechanisms as a means of ensuring constitutional accountability recognizes . . . the reality of privatization, and the need to develop a constitutional analysis that better addresses the constitutional challenges that privatization poses.

The private delegation decisions are additionally instructive in suggesting that express grants of power to private hands merit special scrutiny. Not all private delegations involve grants of government power, and evasion of constitutional constraints can occur through tacit authorizations as well as express ones. Moreover, some of the most vexing state action cases are poorly described in private delegation terms. Yet the recent privatization examples and prior state action cases demonstrate the importance of express private delegations in statutes, contracts, and the like as means of transferring government power to private hands. Indeed, the frequent presence of express delegations of power to private entities is no accident, and instead indicates that most instances of privatization in this country represent moves to private implementation of government programs rather than government disengagement from an activity altogether. As a result, focusing on private delegations is an appropriate method for identifying situations that may present significant threats to constitutional accountability.

Existing private delegation decisions thus provide the seeds for a new constitutional analysis of privatization. Like these decisions, the proposed analysis is structural in focus, in that it targets the inquiry on delegations of power to private actors and on the overall system of constraints to which these delegations are subject. Unlike the prior decisions, however, this analysis generally accepts that private actors can exercise the type of power at issue; it focuses on assessing the impact the private delegation has on the constitutional imperative of accountability. The central concern is determining whether a private delegation is structured so as to ensure that private exercises of government power do not violate constitutional requirements. Constitutional accountability thus has primacy, as the government is prohibited from delegating government power in ways that are not sufficiently constrained. But the government also has flexibility in structuring its relationships with its private partners because various mechanisms exist by which constitutional constraints can be enforced.

Notwithstanding its emphasis on adequate regulatory protections, two central features make this analysis at its core a constitutional inquiry. First, the requirement that adequate protections be extended to privatized programs rests ultimately not on political will, but instead on the Constitution itself, which imposes fundamental limits on how government can act. Second, the Constitution provides the specific substantive content of the protections that regulatory mechanisms must afford, as well as the procedural requirement that these mechanisms must allow for individual enforcement. Even when involvement of private entities affects the substance of constitutional rights, the determination of what protections must be afforded remains a constitutional one: It turns on assessing specifically the constitutional interests at stake, as opposed to simply determining which protections will yield the best policy outcome, “all-things-considered.”
Yet under this approach constitutional law will function differently from the standard court-centered image of constitutional adjudication, where the courts bear primary responsibility for enforcing constitutional requirements. Here the courts provide the constitutional baseline, but the task of translating that baseline into practice falls to the elected branches of government. Moreover, this approach transforms the constitutional right being asserted. The fundamental claim at issue becomes not that the private delegate exercised its powers in ways that violated the plaintiff’s constitutional rights, but rather that the private delegation itself violates the Constitution because it fails to ensure a sufficient level of constitutional accountability. This latter constitutional claim finds a textual home in the Due Process Clauses because allowing exercises of government power outside of constitutional constraints is a violation of the clauses’ prohibition on arbitrary government action. But it also can be seen as rooted not in a particular constitutional text, but in the structural Constitution, because—like separation of powers or federalism—the principle of constitutional accountability that this claim embodies is one of the basic structural postulates of our constitutional system.

Adopting a constitutional analysis that focuses on the overall system of constraints applicable to an exercise of government power offers several advantages. A central benefit is, of course, the potential for better safeguarding of both government regulatory prerogatives and constitutional accountability. In addition, adopting this focus avoids the all-or-nothing character of current state action doctrine. Rather than reducing the inquiry to a single question—whether a particular private entity is a state actor—it pursues a multi-step approach that neither starts nor stops with identifying private exercises of government power. Private delegations are singled out for enhanced constitutional scrutiny because of the possibility, not the actuality, that they involve grants of government power; even if they do, a private delegate’s constitutional immunity is not problematic provided that constitutional constraints on government power are otherwise preserved. A third, and particularly important, advantage is that this analysis aligns the government’s interests with the goal of preserving constitutional accountability. Instead of creating perverse incentives for government to delegate power without strings attached or to control private actors at one step removed, this approach gives government an incentive to ensure that its private delegations are adequately constrained.

B. Sketching the Contours of a Private Delegation Analysis

Such a private delegation analysis holds promise in theory as a remedy for the flaws of current state action doctrine. But as the Supreme Court’s prior failure to develop a coherent private delegation doctrine indicates, the real question is whether this program can work in practice. Three elements of the proposed analysis are central to its success: (1) identifying constitutionally troublesome private delegations; (2) assessing whether such delegations are adequately structured to ensure constitutional accountability; and (3) remediying unconstitutional delegations. A fourth important issue concerns the relationship between the proposed analysis and current state action doctrine and the likelihood of the former’s adoption.
1. Identifying Which Private Delegations Matter: The Centrality of Acting on the Government’s Behalf and Agency

One major obstacle to developing the proposed private delegation inquiry is the extremely broad scope of what could qualify as a private delegation, even limiting the field to express delegations. Licenses, corporate charters, and rights of property and contract, such as the warehouser’s self-help remedy at issue in Flagg Bros. v. Brooks (1978), are “delegations” of state power to private actors as much as government contracts with private service providers or statutes authorizing private regulation. But if all law is understood to represent a private delegation of government power, the effect of targeting private delegations for special scrutiny would be the same as casting aside any constitutional distinction between public and private—something, again, the Court is plainly unwilling to do.

Yet it is a mistake to conclude that for constitutional purposes no meaningful or legitimate basis exists upon which to differentiate among private delegations. A central characteristic of much government privatization is that private delegates are granted powers not simply for their own advantage, but rather to enable them to act—and more specifically, to interact with third parties—on the government’s behalf. . . . [T]his characteristic of acting on behalf of government is what makes these private delegations particularly threatening to the principle of constitutionally-constrained government. By effectively stepping into the government’s shoes in its dealings with third parties, private entities are more likely to have access to powers that are distinctly governmental. These include not simply the ability to exert coercive powers on a nonconsensual basis, but also control over access to government resources and government programs. Particularly when privatization occurs in contexts where program participants or applicants have a great need for the government benefits and services at issue, private entities’ roles in implementing government programs may significantly augment their powers over others and enhance their ability to cause harm. Moreover, this enhancement of private powers is often undertaken in lieu of direct government involvement, which suggests that governments may be evading constitutional requirements simply by changes in form rather than substance. Finally, private entities are often the face of government for participants in government programs, and thus such private actors’ freedom from constitutional controls risks eroding popular belief in constitutionally-constrained government.

Notwithstanding the clear importance to constitutional accountability of a private entity acting on the government’s behalf, the Court’s state action cases have largely ignored this feature. Private law, however, has long recognized the special dangers of relationships in which one party is empowered to act on another’s behalf and addressed these relationships under the law of agency. “Agency is the fiduciary relationship that arises when one person (‘a principal’) manifests assent to another person (‘an agent’) that the agent shall act on the principal’s behalf and subject to the principal’s control, and the agent manifests assent or otherwise consents so to act.” Agency law makes clear that individuals can be acting on another’s behalf even when their actions also advance their own interests; indeed, such dual motives are endemic to agency relationships. But the bare fact that an individual’s pursuit of her own interest benefits another is insufficient to
create an agency relationship; instead, both principal and agent must assent to the agent’s acting on the principal’s behalf. A further essential element of agency is ongoing principal control: “A relationship is not one of agency within the common-law definition unless . . . the principal has the right throughout the duration of the relationship to control the agent’s acts.”

Agency offers a useful means for identifying private delegations that are especially threatening to constitutional accountability and therefore merit greater scrutiny. With its focus on identifying acts on another’s behalf, agency targets precisely the characteristic of private delegations that makes them constitutionally troublesome. Moreover, like the constitutional inquiry into the status of private delegates, a central concern of agency is balancing interests of third parties with those of principals and agents, and much of agency law is similarly focused on determining when it is fair to attribute responsibility for the actions of an agent to a principal. Agency law’s requirements of principal control and manifestations of assent also resonate from a constitutional accountability perspective. The dangers of constitutional evasion are particularly great when the government retains ability to control its delegates; in addition, the presence of assent eases concerns regarding unfairly intruding on regulatory choices or individual autonomy, because both the government and the delegate are able to foresee the likely constitutional repercussions of their relationship.

Of particular importance is agency law’s approach to principal control. At first glance, agency’s insistence on principal control appears problematically akin to current doctrine’s insistence on identifying government involvement—either participation in specific private acts or more general pervasive entwinement in the private entity—before finding state action. Significantly, however, agency law has adopted fairly broad tests for determining when the triggering condition of principal control is met. What matters from an agency perspective is that a principal has the power to control an agent’s actions, not that it actually controls a specific act. Moreover, principal control is compatible with independent decisionmaking and discretion, and agents may enjoy extraordinarily wide authority to make decisions on a principal’s behalf. In contrast to current state action doctrine, governments are not able to escape findings of agency simply by avoiding involvement in day-to-day decisionmaking; instead, they would have to largely cede any right to control a private entity’s actions. Of course, it remains possible that a government would cede such right of control or would restructure its programs—for example by repealing statutory provisions requiring the government to provide certain benefits—in order to prevent a court from finding that private entities were acting as agents on its behalf. But significant political (and legal) obstacles will likely exist to the government’s giving out funds or regulatory power to private actors with so few strings attached.

As a result, focusing on agency offers a means of identifying at least some constitutionally troubling private delegations while still satisfying the demand for a constitutional distinction between public and private. Blum and Rendell-Baker are instructive on the difference resulting from an agency analysis compared to current state action doctrine. While in both decisions the Court held that state action was lacking, the underlying facts plainly support finding that the nursing home and school were government agents: Both provided services to third-party beneficiaries that the
government was statutorily required to make available; moreover, the provision of these services was subsidized by the government and undertaken pursuant to express government authorization. Perhaps most importantly, while not mandating the specific decisions in question, in both cases the government exercised ongoing control over provision of these services through regulations and extensive agency oversight. Like the doctor providing medical services to prisoners in *West*, the nursing home and school were, in essence, independent contractors. Notably, independent contractors can qualify as “agents” for many purposes, even though they enjoy greater freedom from principal control than do employees. State action doctrine, with its emphasis on actual exercises of control, ends up privileging the independent contractor relationship for constitutional purposes compared to the employee relationship, but from an agency perspective little reason exists for drawing such a categorical line.

Indeed, many instances of privatization are fairly clear-cut instances of agency relationships. Private intermediaries processing claims for Medicare, for example, are statutorily designated government agents and have been so recognized. In *Lebron v. National Railroad Passenger Corp.* (1995), the Court indicated that when “the government creates a corporation by special law, for the furtherance of governmental objectives, and retains for itself permanent authority to appoint a majority of the [corporation’s] directors,” the corporation is an agency of government for constitutional purposes. *Brentwood Academy* could have been decided on a similar basis, with public school officials’ dominance of the Association (particularly given the State’s express delegation of oversight responsibility for interscholastic athletics in the past and the Association’s assumption of such responsibility) demonstrating that the Association was acting on behalf of the state and subject to its control.

There are also instances of privatization in which no basis exists to conclude that private delegates are acting on the government’s behalf and therefore agency is lacking. Assent to act on the principal’s behalf appears generally absent in contexts when the government does no more than purchase goods and services for its own use from a private entity; here, the private provider is not interacting with third parties, even its employees, as the government’s representative or on the government’s behalf. *Sullivan* suggests another example. There, the private insurance companies challenging workers’ claims for benefits cannot be described as acting on the government’s behalf, except in the attenuated sense that having private employers provide such insurance served the government’s policy goals of ensuring compensation for all work-related injuries. Except in its role as employer, the government bore no statutory responsibility to provide medical services and gained no financial benefit when medical services were ruled unnecessary. Particularly viewed against the history of workers’ compensation systems, Pennsylvania’s authorizing insurers to deny claim payment until medical necessity is determined seems more a realignment of the respective rights of workers and employers in the employers’ favor than a delegation creating an agency relationship. To the extent that the private insurers were functioning as agents at all, they were serving as agents of employers (who bore statutory liability for workers’ compensation benefits) and not of the government.
In a number of privatized contexts, however, the presence or lack of an agency relationship will be harder to determine. Is a corporation always a government agent when the government appoints a majority of its directors, even if the government makes no effort to exercise control over the corporation’s operations? What if such board and operational control is lacking, but the government charters the corporation and grants it sole regulatory authority over a sphere of activity with the aim of advancing the national interest? Does a statutory provision mandating employee drug testing for public safety purposes transform an employer who engages in such drug testing into a government agent vis-à-vis its employees? Government grants often may create agency relationships, because the government frequently conditions award of the grant on performance of a specific project and closely regulates grantee behavior. Where this is the case, use of grants rather than direct reimbursement for services is an administrative detail that seems to lack wider import. But grants can also represent government efforts to foster independent private action, and distinguishing between the two contexts no doubt will prove difficult.

Moreover, while a useful starting point, agency cannot be an exclusive test for identifying all constitutionally troublesome private delegations. The ultimate issue from a constitutional perspective is whether private entities wield government power, not whether they qualify as government agents. Instances will exist where private actors are delegated what appear to be government powers, but the conditions of agency will be lacking. *Georgia v. McCollum* (1992) is a case on point. There, the Court held that a private defendant’s use of peremptory challenges in juror selection constituted state action, in part because the defendant was helping to select the members of an official government body. In this sense, the defendant could be seen as acting on the government’s behalf; plainly, however, the defendant used the peremptories for his own advantage and his actions were constitutionally protected against prosecutorial control, negating an agency relationship. In other contexts, agency’s requirement of ongoing principal control may preclude finding an agency relationship, even though private entities are acting on the government’s behalf and exercising government powers.

Thus, the existence of an agency relationship should be a sufficient but not a necessary condition for subjecting private delegations to greater scrutiny. It will always remain necessary to investigate whether the delegation represents such a clear grant of government power or power to act on the government’s behalf that further scrutiny is warranted. Yet even though some or many hard cases will remain and it cannot be an exclusive test, agency is an important criterion in singling out which delegations merit special scrutiny. Over time greater clarity will likely develop, and focusing on agency relationships serves to narrow the range of hard cases. In addition, room exists to err on the side of finding agency because, unlike state action determinations, such a finding does not inexorably result in direct application of constitutional constraints on the delegates’ actions. Instead, it triggers an investigation into whether constitutional accountability concerns are adequately addressed.
2. Structuring Private Delegations to Ensure Constitutional Accountability

Determining when private delegations are adequately structured to ensure constitutional accountability represents the second key element in the proposed private delegation analysis. Private delegations that do not create agency relationships or involve clear grants of government power do not require further scrutiny. Such delegations may provide private entities with significant powers, but given the constitutional public-private divide, they do not raise constitutional accountability concerns.

This leaves, however, the many private delegations that do have these characteristics, creating the danger that private delegates are stepping into the shoes of government and wielding government power. When are such private delegations adequately structured to preserve constitutional accountability notwithstanding that the private delegates involved are exempt from the Constitution’s purview? One central requirement, derived from the definition of constitutional accountability, is that a mechanism must exist by which individuals can enforce constitutional constraints on government power, ultimately in court, even if they cannot assert constitutional rights directly against private delegates. This requirement, in turn, suggests two ways in which private delegations can be structured to meet constitutional accountability concerns. Under the first, alternative mechanisms exist through which individuals can assert constitutional constraints (or adequate surrogates) on exercises of government power even though private delegates wielding government power are immune from direct constitutional scrutiny. Under the second, private delegations are significantly limited, so that the private delegates do not in fact wield government power—or if they do, their exercises of this power are constitutionally insignificant. Both of these approaches allow room for a court to conclude that the content of constitutional protections may differ in the context of private delegations in comparison to when government officials act alone.

a. Alternative Mechanisms to Enforce Constitutional Constraints

The most obvious alternative mechanism for enforcing constitutional constraints is government supervision of private decisionmaking. Government supervision is, of course, the structural mechanism emphasized in existing private delegation decisions that reject claims that private delegations contravene separation of powers and due process requirements. Government supervision also prefigured in two of the Supreme Court’s recent encounters with privatization, Correctional Services Corp. v. Malesko (2001) and Richardson v. McKnight (1997), both of which involved private prisons. In Malesko, the Court emphasized that inmates in private facilities could access the Bureau of Prisons’ administrative complaint system and from there obtain judicial review in denying a Bivens action for constitutional violations against the private prison operators. Meanwhile, in Richardson the Court held that individual guards at private prisons generally do not enjoy qualified immunity against constitutional claims under § 1983, but the Court expressly reserved the question of whether qualified immunity would be available to private individuals “acting under close official supervision.”

Government supervision is a particularly important mechanism for ensuring accountability in privatization contexts. The government’s regulatory and contractual
powers, as well as its administrative resources and expertise, put it in the best position to control private delegates’ behavior. In order to meet the demands of constitutional accountability, however, government supervision must take the form of a complaint or appeals system akin to that in *Grijalva* or *Malesko*, through which affected individuals can challenge specific private decisions, policies, and procedures. Such a review system not only provides a direct administrative check against private delegate abuses, but also provides the requisite basis for a subsequent constitutional challenge in court. If the government denies relief, individuals then can bring constitutional challenges to the government’s own decisions—for example, alleging that the government violated due process by sanctioning inadequate private decisionmaking procedures or by affirming a private decision made on racially discriminatory grounds. On occasion, preserving constitutional accountability may require additional safeguards, such as a means for immediate review where urgently needed assistance is at stake, or de novo review where the government delegates enforcement of regulatory requirements to private hands.

One advantage of using government oversight to ensure constitutional accountability is that government officials are not limited to rectifying constitutional violations, but can also use their oversight to enforce regulatory requirements and contractual promises, and in other ways improve the effectiveness and quality of privatized programs. While governments rarely can be required to undertake such enforcement, the necessity of providing an individual appeals mechanism gives them an incentive to do so. This mechanism supplies governments with needed information about the performance of their private delegates, and also allows governments to review the quality of services in individual cases without significant administrative cost. Moreover, enforcing regulatory and contractual requirements regarding delegates’ behavior will lessen the danger that private delegates will act in constitutionally prohibited ways, for which the government might subsequently be found liable. On the flip side, however, governments are not left with much flexibility regarding how they conduct oversight of their private delegates. Since a large part of the motivation behind privatization is the government’s desire to avoid involvement in implementation decisions, holding that private delegations can be rendered constitutional by such involvement seems to offer little solace.

Here, an important point to emphasize is that unlike government proper, not all actions by private government agents trigger constitutional consequences; instead, constitutional requirements attach only when private entities are wielding powers derived from their agency relationship with the government. As a result, where a private delegate’s actions are ones that it otherwise had authority to undertake and which relate tangentially to the responsibilities it performs for the government, more minimal government oversight should suffice. *Rendell-Baker* is illustrative. There, the private school fired teachers who had criticized aspects of the school’s operations, and the teachers challenged their terminations on constitutional grounds. While the school qualified as a government agent, given its responsibility to provide special needs education on behalf of the local school committees, it would have had the same power over its teachers absent the committees’ delegation of this responsibility. What makes the teachers’ firing nonetheless seem constitutionally troubling is the importance of the teachers’ ability to criticize the school as a protection against the school abusing its delegated powers over students. Yet this constitutional concern seems adequately addressed if some means exists by which
teachers and parents can bring issues regarding the school’s operation to the
government’s attention. In *Rendell-Baker*, such a means was present in the guise of
informal oversight by the state.

A second point is that close government supervision of private delegates is not the only
alternative mechanism by which to ensure constitutional accountability. The government
instead could create separate decisionmaking tracks, so that affected individuals obtain a
completely independent decision from the government in place of the private delegate’s
ruling. Another option is for the government to forego being involved in oversight and
instead to make private delegates directly subject to substantive and procedural
requirements that represent adequate surrogates for constitutional protections. Such
requirements could be based in federal or state law (depending on the level of
government where they originate) and imposed by statute, regulation, or contract. Title
VI’s prohibition on recipients of federal funds engaging in race discrimination is a
statutory example of this approach. A third variation is for the government to impose a
general requirement that private delegates assume constitutional responsibilities, rather
than detail the substantive and procedural constraints to which delegates must adhere. In
order for these latter two approaches to satisfy constitutional accountability concerns, a
means by which individuals could enforce these requirements against private delegates,
such as an express right of action, must exist.

Finally, state law rights not specific to privatization contexts may suffice to meet
constitutional accountability requirements. In some situations, private remedies, such as a
tort action for negligence or a suit under generally applicable civil rights statutes, may
provide equal or greater protection than is available through constitutional litigation.
State law may also provide a basis for public law actions that force the government to
remedy abuses by its private partners. In *Logiodice v. Trustees of Maine Central Institute*
(1st Cir. 2002), the First Circuit concluded that under Maine law, school districts are
statutorily obligated to provide students with a free education. That a school district
contracted with a private school to perform this function did not relieve the district of its
statutory duty, and thus a student wrongly expelled could sue in state court to compel the
district to provide the free education to which she was entitled. One point to note here is
that the emphasis on acting on the government’s behalf and agency may limit the
availability of some private remedies, as this emphasis may expand the scope of private
actors entitled to government contractor immunity.

In assessing the adequacy of these alternative mechanisms, courts will by necessity have
to address the question of whether the involvement of private delegates changes the
substance of constitutional protections. The principle of constitutional accountability
might at first suggest that the presence of a private delegation should have no such effect,
so as not to create an incentive to delegate government power. Under such an approach,
to be deemed constitutionally adequate, an alternative accountability mechanism would
have to impose constraints at least as great as those the Constitution imposes on
government. Weighing against this conclusion are the constitutional protections for
private associational autonomy and other freedoms that may be unduly infringed if
private delegates are found to be subject to the identical constitutional prohibitions as the
government. Thus, taking full account of constitutional concerns requires courts to balance the conflicting constitutional values at stake.

Importantly, however, the question of whether private delegates’ involvement affects the substance of asserted constitutional constraints comes framed by a court’s antecedent determination that the private delegate is in an agency relationship with the government, and therefore is likely to be wielding government power. This means that such judicial balancing of constitutional rights and private autonomy concerns is necessarily interstitial: It does not extend to activities that fall on the private side of the public-private divide; moreover, the content of constitutional constraints on government represents the baseline from which any deviations must be justified. As a result, it is less threatening to the political branches’ regulatory prerogatives than the wholesale balancing associated with casting aside the state action doctrine altogether. Indeed, balancing in this form is a common feature of ordinary constitutional adjudication, where the specific substance of constitutional rights frequently varies by context.

b. Limits on Private Delegations

Governments can also ensure constitutional accountability by limiting the powers they delegate. While an agency relationship between private delegates and the government increases the risk that private delegates are exercising government power, this result is not inevitable. On some occasions, acting on the government’s behalf does not give private delegates enhanced power over others. One such situation is where there is a large pool of private delegates who compete with one another and wield the same authority on the government’s behalf, with the result that no delegate exercises monopoly or even quasi-monopoly control over program participants or access to government benefits. This renders it unlikely that any particular delegate’s powers will be enhanced significantly by the delegations in question, or that the government will be able to exert control over privatized programs without actively intervening in specific decisions. Medicare’s fee-for-service program, under which every physician licensed by the state is eligible to provide government-subsidized care, represents one example. The Food Stamp and Section 8 housing voucher programs are others.

Part of what minimizes accountability concerns in these contexts is the role played by individual choice. The authority to make decisions on behalf of government is exercised by beneficiaries who are directly affected by the decisions and who, in principle at least, can protect themselves against abuse by choosing different providers. If anything, the private delegates most empowered by this type of delegation are the beneficiaries themselves; and the fact that beneficiaries make decisions primarily for themselves rather than for others minimizes the danger that the delegates’ expanded powers represent an exercise of government authority. Seen in this light, this structural device is not limited to broad delegations, but is also present in some self-regulation contexts where affected individuals directly participate in setting the content of binding norms.

A key requirement, however, is that individual choice and right of participation must be real. Thus, if the nature of the services at issue makes it difficult to switch providers once services have begun, then a broad pool of delegates’ and beneficiaries’ power of choice
may not protect against private delegates wielding government power. Given their relatively closed environments and the trauma that the elderly and vulnerable experience when forced to transfer, nursing homes can exercise tremendous power over residents. Hence, the fact that Medicaid beneficiaries can choose from a broad array of private homes in deciding where to reside does not remove the constitutional accountability problem in *Blum*, where homes were authorized to determine residents’ eligibility for different types of care. In addition, for meaningful choice, individuals must be informed of the availability of alternative providers and the procedures by which they can switch providers. The outer range of constitutionally-required information is unclear. Arguably, where the services involved are of particular importance and difficult to assess, ensuring a meaningful choice requires that the government provide detailed information on the services provided by different providers and perhaps offer additional assistance to beneficiaries in making their selections. Yet such a requirement is at least superficially hard to square with the Court’s refusal to impose affirmative assistance obligations in many instances where the government is acting directly against an individual’s interests.

A final alternative is to design the delegation in such a way that private entities’ exercise of their delegated authority necessarily comports with substantive constitutional requirements. *Currin* provides an example of this approach. There, tobacco growers had the sole power to determine whether tobacco standards promulgated by the Secretary of Agriculture would apply in a particular area. Thus, the delegation was not one where government retained the power to review private decisionmaking, nor one where all affected individuals participated. Yet the growers could determine only whether the standards would become effective; they could determine neither the content of the standards, nor how the standards would apply in particular cases. In rejecting the claim that the statute was a private delegation of legislative power, the Court emphasized the growers’ limited role, but the point applies more broadly. The tobacco growers’ power was so circumscribed that its exercise, even for purely biased and self-interested ends, would not violate the Constitution. Indeed, self-interested exercise was the very point. This was simply an example of Congress regulating to improve the growers’ economic position, as it had power to do, but letting the growers determine whether such regulation really was to their advantage.

How far this final option extends is unclear, and existing precedent seems to suggest a narrow scope. Although *Sunshine Anthracite* did not expressly hold that government oversight is a constitutional imperative when some private actors are delegated power to promulgate regulatory standards binding on others, this conclusion follows from the Court’s reasoning. Similarly, professional norms impose standards often exceeding the scope of constitutional prohibitions. Yet, in *West*, the Court rejected the claim that delegating government power to professionals obviated the need for individuals to be able to assert constitutional constraints. On the other hand, these decisions focused on the question of whether the private actors involved were wielding legislative power or were state actors, not whether the move to privatized contexts altered the applicable substantive content of constitutional guarantees. Once the fact that private delegates are wielding government power is acknowledged, this latter question moves to the fore.
3. Remedying Unconstitutional Private Delegations

The discussion so far has sketched most of the contours of the proposed private delegation analysis. An additional important question that remains, however, concerns how courts should remedy delegations held unconstitutional.

a. Invalidation

In prior private delegation cases and analogous contexts, the court employed a remedy of granting declaratory or injunctive relief in holding a delegation valid. . . . It would then be up to the government to add the required accountability mechanisms or otherwise refashion the delegation to be constitutional. Not only is this remedial route the most logical response to an unconstitutional delegation, but more importantly, it accords with the basic claim underlying the proposed analysis: that the political branches of government bear primary responsibility for structuring private delegations to ensure constitutional accountability. It is, however, subject to two potentially significant criticisms. First, invalidating unconstitutional private delegations could be very disruptive, particularly given the heavy reliance on private actors in social welfare programs, where beneficiaries are often in urgent need for services and continuous care relationships are particularly important. Second, this remedy would fail to rectify the harm an individual suffers as a result of an inadequately constrained private delegation; for example, simply invalidating a welfare program’s reliance on private case managers does not make whole a welfare recipient who was dropped from the rolls or denied an opportunity to participate in a job training program because of her race.

b. Direct Application of Constitutional Constraints to Private Delegates

An alternative approach would be for a court to remedy an unconstitutional private delegation by applying constitutional constraints directly to the private delegates at issue. Governments then could continue to rely on their private delegates, and courts could ensure full relief by proceeding to consider the merits of an individual’s constitutional claims if they find a private delegation to be inadequately structured. This option has the further advantage of being most in keeping with current practice. Indeed, the effect of this remedial option would be to transform the private delegation approach into a version of state action analysis, one that—unlike current doctrine—allows for contingent ascriptions of state action. On this account, a private delegate’s state actor status would depend upon the absence of adequate alternative accountability mechanisms. Such greater compatibility with existing law is important, because it may dramatically increase the likelihood that the proposed private delegation analysis will be judicially adopted.

These benefits come at a significant cost, however. Such a state action remedy once again collapses the problem of how to constrain government power into the question of whether private actors should be subject to constitutional requirements, thereby undermining the proposed analysis’ structural focus on grants of power and overall constraints. Moreover, this approach would erode the government’s incentives to structure its private delegations adequately from the outset. Equally problematic are the potential effects on judicial review. Courts are more likely to insist on exact surrogates for applicable constitutional
constraints when imposing such constraints on private actors represents the immediate alternative; as a result, they will be less willing to accept accountability mechanisms that are different in structure as constitutionally adequate alternatives. At the same time, the opposite danger also exists: that courts will adopt an excessively narrow account of when private delegations trigger heightened scrutiny to avoid subjecting private individuals and entities to constitutional requirements. Perhaps the latter result is likely regardless, given the extent to which the current Court seems willing to ignore the reality of privatized government in state action cases. But its probability is only enhanced by relying on a remedy that requires courts to impose liability for constitutional defects on private actors rather than on government.

c. Invalidation with Direct Application if No Adequate Alternative Exists

On balance, the route of invalidation in a suit against the government appears generally the better remedial option. The remedial stage is where the equitable powers of the courts are at their broadest and thus courts should not have difficulty fashioning interim relief that minimizes disruption or ensures that injured individuals are made whole. Courts should be able to hold the government responsible to rectify harms caused by unconstitutional actions of private delegates, not because those delegates are state actors, but rather because the government itself violated the plaintiffs’ constitutional rights by delegating government power without adequate safeguards. Courts also could respond to a delegation’s potential unconstitutionality by reading the grant of delegated powers narrowly, a remedy much more commonly used to address delegation concerns than outright invalidation. Courts also can invoke separability doctrine.

On the other hand, the rejection of direct application of constitutional constraints to private actors should not be absolute. The barrier is prudential only, assuming that the private delegates are defendants in the case. Once a court has found an agency relationship indicating private exercise of government power, no substantive constitutional barrier prohibits subjecting private actors to constitutional requirements. In some contexts, an adequate alternative to direct constitutional application will not exist, whether because of the pervasiveness and potency of the delegate’s powers or because other means of ensuring accountability are not readily available. One example of the latter situation is a defendant’s exercise of peremptory challenges, which the Sixth Amendment protects from government supervision other than by a court. The choice then becomes whether to apply constitutional constraints directly to private delegates in such contexts, or to prohibit these private delegations entirely. While either of these options would satisfy constitutional accountability concerns, the proposed analysis’ emphasis on respecting the political branches’ regulatory freedom counsels for direct constitutional application rather than prohibition.

In other contexts, the government may signal its willingness to have its delegates bound by constitutional requirements, for example, by statute or in its contracts with private entities. Where the government has so provided, the regulatory prerogatives argument against direct application of constitutional norms is once again quite weak. The real question is instead one of forum: Should plaintiffs be forced to sue under these stipulations, which often would be based in state law and thus would limit them to state
court? Or should plaintiffs also have access to federal courts based on direct constitutional claims, even though these stipulations arguably cure any constitutional accountability problem with the underlying delegation? Given that the federal Constitution is the ultimate measure of the rights being asserted, requiring plaintiffs to sue again in state court appears hard to justify. Hence, here too, the appropriate course will often be to hold that constitutional rights apply directly to the private entities involved, thereby ensuring access to federal courts under existing jurisdictional statutes.

4. The Effect of the Proposed Analysis on Existing State Action and Private Delegation Doctrine

A final issue concerns how the proposed private delegation analysis would interact with existing constitutional approaches to privatization. This analysis is derived from existing private delegation decisions and requires little alteration of that doctrine. But it would force a marked departure from how courts address what are now seen as state action problems.

Most notably, under the proposed analysis, the government’s substantial involvement in specific private acts or more general “pervasive entwinement” with a private delegate would not subject the delegate to direct constitutional liability. On the contrary, close government involvement will often serve to defeat constitutional objections. While such involvement supports finding that the private delegate is a government agent and is exercising government power, it increases the probability that suits against the government alone are available as alternative means to ensure constitutional accountability. The private delegation approach is more compatible with the present public function test; in fact, the category of private delegations found to represent clear grants of government power could be defined alternatively as cases where a private delegate is exercising powers that traditionally and exclusively were reserved for the state. But here as well the logic of the proposed analysis would force alterations. No reason exists for subjecting private delegates to constitutional constraints simply because they are performing a public function if constitutional accountability is otherwise achievable. Often, however, application of the proposed analysis would not change the result in public function cases, because the types of power at issue in those contexts tend to be such that no other mechanism adequately ensures constitutional accountability.

The inconsistency of the proposed private delegation analysis and current state action doctrine, of course, makes it unlikely that the Court will adopt the proposed approach. From a realist perspective, the odds of judicial adoption are further diminished by the fact that the proposed analysis will subject many public-private relationships assumed to be unproblematic under state action doctrine to a much more searching review, with some no doubt being invalidated. Yet as privatization expands, courts increasingly must confront the transparent inadequacies of current state action doctrine. Judges will be in the uncomfortable position of choosing between constitutionalizing private action and thus displacing the political branches’ regulation of government-private relationships, or allowing substantial exercises of government power to evade constitutional scrutiny. Faced with such an unpalatable choice, courts may become more accepting of efforts to
fundamentally rework state action analysis, and more willing to impose obligations to preserve constitutional accountability on governments as the price of privatization.

If so, the private delegation approach has notable advantages compared to [other] proposals for reforming state action doctrine. . . . Determining whether an agency relationship exists may at times be difficult, but, even so, agency law provides more principled and independent guidance for courts than amorphous tests such as “public function” or “pervasive entwinement.” While offering a more formalized inquiry, the proposed analysis still allows courts to take account of the realities of government-private relationships; these realities will factor into a court’s determinations of whether agency is present and whether the private delegations at issue are adequately structured to preserve constitutional accountability. Unlike prior reform proposals, the private delegation approach avoids unnecessary restrictions of the political branches’ regulatory role by emphasizing alternative accountability mechanisms and foregoing direct application of constitutional constraints to private actors.

Perhaps most significantly, notwithstanding its inconsistency with current state action doctrine, the proposed analysis accords with the basic constitutional propositions on which that doctrine is based. One such proposition is acceptance of a public-private divide for constitutional purposes. True, the proposed private delegation approach draws this line differently from current doctrine. But the waxing and waning of state action doctrine over the second half of the twentieth century makes clear that there is more than one way of construing the meaning of government power. What the public-private divide requires is simply a recognition that the sphere of the Constitution’s application is limited, so that certain actions are not subject to constitutional scrutiny. But it does not prohibit drawing the appropriate dividing line based upon a measured assessment of how best to balance constitutional accountability and government regulatory freedom.

Similarly, no conflict exists between the proposed analysis and the principle that the Constitution applies to all state action, even though private exercises of government power are immune from direct constitutional scrutiny. Under the proposed analysis, constitutional constraints on government power are preserved, albeit through a different means. Private delegates’ general immunity from direct constitutional scrutiny can also be justified on the grounds that when the state has imposed adequate constraints on its private delegates’ exercises of government power, such exercises should not be deemed to constitute state action for constitutional purposes.

The proposed analysis is also compatible with the principle that, absent exceptional circumstances, government owes no duty to protect individuals from independent private harm. True, the effect will often be to require the government to provide protections against private abuse; indeed, one central advantage of the private delegation approach is that it encourages the government to undertake greater supervision of its private delegates. Critically, however, private delegations do not represent situations in which the government has failed to intervene to address harms caused by private actors, but rather ones where the government has affirmatively acted by delegating decisionmaking authority over government programs to private hands. Thus, holding that government bears responsibility for ensuring private delegations are adequately constrained is simply demanding that the government itself act in accordance with constitutional
requirements—in this case, with the principle of constitutional accountability. Here again, the structural emphasis of the proposed inquiry is useful because it directs attention to the government’s own decisions regarding the design of private delegations, conduct for which the government is clearly liable.

C. Application of the Proposed Private Delegation Analysis

To summarize, the proposed private delegation analysis would proceed as follows: An individual injured or likely to be injured by a private delegate’s action would bring suit against the government (and perhaps the private delegate itself), arguing that the delegation authorizing the private action is unconstitutional because it is inadequately structured to ensure that private exercises of government power adhere to constitutional requirements. In assessing the merits of this claim, the court would first determine whether the delegation creates an agency relationship between the private actor and the government. If not, and if the private delegation does not otherwise involve a clear grant of government power or power to act on the government’s behalf, the plaintiff would lose on the merits.

Where the court finds such an agency relationship, it would then proceed to assess whether the delegation is adequately structured to preserve constitutional accountability. This requirement is met where some mechanism exists by which the individual can enforce constitutional limits on private exercises of government power, such as a statutory surrogate for a constitutional claim or the ability to assert actual constitutional claims against the government. The government becomes a potential target for constitutional challenge when it either sanctions the private delegate’s specific actions on administrative review or makes an independent determination as an alternative to private decisionmaking. Other means of satisfying constitutional accountability demands would be to structure the delegation so that no private delegate exercises monopolistic or quasi-monopolistic control over access to government benefits, or to delegate only very limited powers. A court may also conclude that a private delegation meets constitutional requirements, even without limitations on private delegates equivalent to those that apply to government, because in a particular context the Constitution imposes lesser substantive constraints on private entities wielding government authority.

As this description suggests, the claim that a private delegation is inadequately structured to preserve constitutional accountability can take the form of a facial challenge. In some cases, however, the inadequacy of alternative accountability mechanisms will only become apparent through practical application; for example, it may not be clear from the face of governing regulations that certain types of claims are excluded from an administrative complaint system. In such as-applied contexts, plaintiffs would combine their claims for relief under existing accountability mechanisms with the alternative argument that if such relief is unavailable, the delegation is unconstitutional.

If a private delegation creates an agency relationship or involves delegation of power to act on the government’s behalf, and is inadequately structured to ensure that government power is ultimately kept within constitutional bounds, the delegation would be unconstitutional. The usual remedy will be for a court to invalidate the delegation and
hold the government responsible for the delegate’s unconstitutional acts. But in rare situations, courts may instead apply constitutional constraints directly to the private delegates involved, because doing so represents the only means of ensuring constitutional accountability.

The major question remaining is what difference the proposed private delegation analysis will make in practice. Some sense of an answer comes from examining the different results in past state action cases, summarized . . . in [the table at the end of this document]. Further indications about the practical effect of the proposed analysis become apparent when applying it to the recent government privatization efforts in the health care, welfare, public education, and prison contexts . . . .

1. Privatization in Health Care: Medicare and Medicaid Managed Care

The statutes, regulations, and contracts authorizing the use of managed care under Medicare and Medicaid represent private delegations that create agency relationships between MCOs and the government. While desirous of making a profit for themselves, the MCOs also plainly act on the government’s behalf: In exchange for a set per-beneficiary payment, the MCOs determine eligibility for and provide medical services that federal and state governments are required to make available to beneficiaries in these programs. The contracts between the government and MCOs, as well as the obvious connection of the MCOs’ services to Medicare and Medicaid, offer ample evidence of principal and agent assent to an agency relationship. Principal control, in turn, is manifested by the government’s stipulations regarding which services must be provided and regulation of MCOs’ operations.

Thus, under the proposed private delegation analysis, these delegations merit enhanced scrutiny to determine if they are adequately structured to ensure constitutional accountability. Regulations limit when beneficiaries can leave an MCO, so that the options of enrolling in a different MCO or returning to fee-for-service (options that may only exist under Medicare) do not satisfy concerns about the MCOs’ powers over beneficiaries. Accreditation requirements are significant protections against poor quality services and denials of medically needed services; however, they are not adequate safeguards from a constitutional accountability perspective because accrediting organizations do not provide oversight of MCO decisions in particular cases and often lack mechanisms by which individuals can challenge accreditation decisions. Private rights of action and rights to independent review of coverage denials may be available under state law, but such rights may not extend to the Medicare and Medicaid contexts given the federal nature of the programs and the presence of administrative systems for review of denials.

As a result, whether these private delegations to MCOs are constitutional turns on the adequacy of these systems and their respective provisions for administrative review. In Grijalva, the court concluded that the existing system for administrative appeals, which closely paralleled the appeals procedures used to challenge denials of coverage under Medicare fee-for-service, failed to satisfy due process in the MCO context where such a denial prevents access to health care, not just access to reimbursement. In particular, the
court concluded that the period within which administrative review had to occur was excessively long and that notices of appeal rights provided by MCOs were inadequate. The due process concerns raised by the court appear quite real, and the conclusion that a system of government review fails to ensure due process would also mean that the system is inadequate as a constraint on a private delegation.

Until this point, therefore, the proposed analysis deviates little from that in *Grijalva*. The difference between the two lies instead at the remedial stage. *Grijalva*’s state action finding would support suits directly against the MCOs on constitutional claims, had it not been vacated post-*Sullivan*. Under the proposed analysis, however, such suits would not be authorized and instead claims of unconstitutional delegations would lead only to relief from claims against the government. Moreover, the content of such relief would be different from that in *Grijalva*. Rather than specifying the substance of the procedures MCOs must follow in detail (including the size of font to be used in notices), courts would simply require the government to adopt adequate constraints or rescind the delegation. It would be up to the government to determine whether to improve its oversight of MCOs or adopt another approach, such as express rights to independent review. In addition, the regulations that the government issued while *Grijalva* was on appeal might well satisfy constitutional accountability requirements, even though they did not comply with the specifications in the district court’s order. Hence, their promulgation would have ended the government’s liability and mooted the case.

Implicit in the foregoing is that the delegations to MCOs raise constitutional accountability concerns only with respect to their treatment of enrolled beneficiaries. MCOs also exercise control over other third parties—specifically beneficiaries seeking to enroll, employees, and participating medical providers. But with respect to these individuals, the delegations are adequately structured to preserve constitutional accountability. Governing statutes and regulations expressly protect the rights of Medicare beneficiaries to enroll in a participating MCO, and the option of fee-for-service in the interim plays a more curative role. Similarly, medical providers’ ability to affiliate with another MCO or offer services through fee-for-service means that no single MCO exercises monopolistic or quasi-monopolistic power over providers’ ability to participate in Medicare. Accordingly, MCOs do not appear to be in an agency relationship with the federal government regarding their other employees.


Little doubt exists that Wisconsin’s delegation to private organizations of responsibility for operating the W-2 welfare program in Milwaukee creates agency relationships between the organizations and the State. Indeed, the W-2 statute makes this clear by defining the entities implementing W-2 as “agen[ies].” Unlike the workers’ compensation system in *Sullivan*, here the obligation to provide benefits lies by statute with the government, which in turn contracted out this responsibility to private agencies, with the state welfare department exercising oversight. The W-2 statute sets out the basic features of the W-2 program that the W-2 agencies are responsible for implementing—including the program’s eligibility and participation requirements, levels of benefits, and types of services available to beneficiaries. Nor can there be much doubt that this
delegation is inadequately structured to preserve constitutional accountability. Although four private W-2 agencies now operate in Milwaukee, each has a monopoly over implementing the program and providing welfare benefits in the particular geographic regions it serves. The W-2 statute imposes few procedural restrictions on the W-2 agencies, and the government’s contracts with W-2 agencies also impose very few restrictions on how their programs operate. The oversight undertaken by the state’s welfare department, while sufficient to support a finding of agency, is insufficient for constitutional purposes; individuals are allowed to petition the department for review of adverse W-2 agency decisions, but the department need grant such review only where the decision found an applicant financially ineligible for aid. A provision in Wisconsin’s public assistance law states that public assistance agencies shall respect the rights of public assistance recipients, including constitutional rights, but it provides no express mechanisms by which recipients can enforce this requirement.

Hence, the delegation in its current form would be unconstitutional. Moreover, adequately restructuring the program to ensure sufficient accountability would appear to require that the government either implement a broader administrative appeals system, or dramatically alter the performance incentive structure of the W-2 program. The current structure gives the W-2 agencies a financial interest in the decisions they make regarding program participants, such that allowing the agencies to make these decisions, at least absent access to independent review, appears to violate due process.

These problems notwithstanding, the W-2 case illustrates how a private delegation inquiry can invest the government with more regulatory freedom than does a state action analysis. It seems likely that the private W-2 agencies would be found to be state actors on the basis of government involvement; as in Grijalva, the W-2 agencies not only provide state-subsidized services, but also implement state-specified welfare eligibility and participation requirements. There is also a solid chance that wholesale administration of a state’s income assistance program would qualify as a public function, particularly given that in most of Wisconsin, the functions of the private W-2 agencies are performed by county social services departments. Accordingly, under current doctrine the W-2 agencies are subject to direct constitutional scrutiny. While the government’s regulatory flexibility is also limited under the private delegation analysis, it at least gives the government the choice of avoiding this result by instituting more comprehensive review.

The opposite situation holds with respect to contracted-out job placement and case management services. Blum and Rendell-Baker indicate that these situations will rarely trigger a state action conclusion. Yet these forms of privatization are instances of private delegations that create agency relationships with the government and that would provoke enhanced scrutiny under the private delegation analysis. Some of these delegations may end up being unconstitutional, but such a result is not preordained. While governments exercise little case-by-case oversight over the actions of such contracted service providers, even a minimal opportunity for review of job placement decisions will likely satisfy constitutional concerns, at least as long as providers have no financial interest in limiting placements and particularly if placement decisions are given to professionals. But under the proposed analysis, courts would be able to assess whether additional procedural safeguards are needed to ensure private providers’ decisions regarding
services comport with constitutional requirements, whereas under current doctrine such private decisionmaking is constitutionally exempt. As a result, courts are able to assess whether additional protections may be required in particular contexts—for example, when a private provider’s determination that a beneficiary has been absent from a placement without good cause results in sanctions or terminations of benefits where constitutional concerns may be more acute.

A more interesting alternative is suggested by the Workforce Investment Act of 1998 (WIA). Included in the WIA is a requirement that states provide individual training accounts as part of their workforce development programs, which allow program participants eligible for training services to choose their providers rather than be assigned. The WIA provides little detail on how these accounts will operate in practice. But assuming the WIA board has contracted with a variety of providers and beneficiaries can truly choose among programs, use of this approach would adequately address constitutional accountability concerns otherwise presented by private providers’ control over government-subsidized employment services.

3. Privatization in Public Education: Charters, EMOs, and Vouchers

A useful starting point for assessing the impact of the proposed analysis on the three moves to greater privatization in public education detailed above—development of charter schools, reliance on private EMOs to manage public schools, and subsidization of private school tuition through vouchers—is to examine how these three initiatives fare under current state action doctrine. Charter schools most likely would be found part of the government for constitutional purposes, given that they are officially denominated public schools, often are created by the state, and operate subject to the state’s direct oversight. At a minimum, these features present a very strong case for a conclusion of state action, either on the basis of extensive government involvement as in *Brentwood Academy*, or on the theory that managing a public school (although not providing educational services) is traditionally the exclusive prerogative of the state. A similar conclusion applies to privately managed public schools and the decisions of EMOs in operating them. On the other hand, *Rendell-Baker* strongly suggests that state action will not exist with regard to private schools participating in a voucher program unless there is evidence the state compelled or encouraged the specific decision at issue.

All three, however, represent delegations where private entities are authorized to act on the government’s behalf, even if not instances of agency relationships. EMOs are perhaps the clearest case for agency: They not only operate acknowledged public institutions and provide education services in exchange for payment from the government, but are most clearly subject to government control through school board supervision and oversight. Charter schools’ frequent exemption from state educational regulations in exchange for performance promises renders government control more limited. But the combination of requirements that the government does impose on charter operation (such as open access mandates) with provisions for chartering agency oversight and charter revocation, may suffice to establish agency; their official denomination as public schools adds an additional basis for finding agency status. On the other hand, while the private schools involved in voucher programs are providing educational services subsidized by the
government and are subject to some government regulation, the relationship between the
government and the private schools appears too attenuated to support a finding of agency.
Nonetheless, private schools participating in voucher programs are fairly described as
acting on the government’s behalf, in that they provide educational services in exchange
for payment by the government, services that the government would otherwise have to
provide directly. Moreover, the schools clearly exercise some control over enrolled
students’ access to a government benefit: a subsidized education.

As a result, the constitutionality of all three private delegations again reduces to the
question of whether the delegations are adequately constrained. Where the government
provides an administrative process under which students and teachers can file complaints
and obtain expeditious independent review of particular actions, such as suspensions or
employment terminations, adequate constraints appear present. But such a complaint
mechanism is unlikely to be available; in the case of EMOs operating public schools,
access to government review mechanisms will likely depend on the terms of the contract
between the EMO and the district, while charter school laws and voucher programs
provide at best for periodic oversight by government agencies. Independently enforceable
statutory or regulatory controls are also unlikely, as states generally impose few
substantive requirements on charter schools or schools participating in voucher programs.
Similarly, the state law remedy relied upon in *Logiodice*—an action against the school
district for failing to meet its statutory duty to provide a free education—will have less
traction where public school alternatives are provided.

This raises the interesting question of whether adequate constraints are provided by the
very system of school choice itself. In principle, such systems of choice should
adequately ensure constitutional accountability because the schools’ powers are
dependent on parent, student, and teacher decisions. A central difficulty with relying on
school choice to preserve constitutional accountability, however, comes from the fact that
all three moves to privatization are much more common in urban school districts marked
by failing public schools. Realistically, therefore, the option of remaining in the public
school system is not available, and ostensible school choice may offer little protection
against private abuse of government power. Yet, *Zelman* indicates, that the Court views
“private choice” in the public education context as meaning formal rather than actual
choice. The Court sustained Cleveland’s voucher program against an Establishment
Clause challenge in part because the program allowed students to attend Cleveland’s
public schools, out-of-district public schools, or private non-sectarian schools—even
while simultaneously acknowledging the abysmal education available through the first
alternative, and notwithstanding that low tuition subsidies made the other two nearly
nonexistent in practice. Holding that constitutional accountability demands use of
vouchers only when the public schools are healthy is also at odds with *Zelman’s*
suggestion that in such contexts a better case can be made that the voucher program is
unconstitutionally motivated by a desire to foster religion. And as a practical matter, that
approach seems to deny the government flexibility to explore privatization precisely
when such flexibility is most needed.

Looking at these instances of public education privatization through a private delegation
analysis thus could have varying effects. It would forestall finding charter schools and

266
Draft — do not distribute!
EMOs directly subject to constitutional constraints as state actors. But, more importantly, it would require states to expand their systems of oversight or impose independently enforceable requirements on charters, EMOs, and private schools participating in voucher programs. On the other hand, the reliance on individual choice to structure these programs suggests that the level of close government supervision required in the Medicare/Medicaid MCO and W-2 contexts should not be required here.

Charter schools and voucher programs are also good examples of how constitutional rights may change with the move to more private contexts. In part, this results from changes in the factual underpinnings that support certain claims; teachers who are hired by charter schools or private schools do not have the same expectation of continued tenure as those hired by public schools governed by civil service or collective bargaining rules and thus may have no procedural due process right to a hearing if fired. Further, it may be that the substance of some constitutional rights changes in order to accommodate the legitimate autonomy and associational interests at stake. To take a contentious example, perhaps religious schools participating in voucher programs should be able to prohibit advocacy of other faiths on school grounds without violating the First Amendment, although no public school could do so. Under current doctrine, this conclusion follows from the lack of state action, but it could also result under the private delegation analysis from recognition of the school’s strong interest in preserving its religious identity. Alternatively, given the government’s interest in preserving private flexibility, perhaps private schools should not be subject to the same due process requirements as public schools, even when similar benefit entitlements are at stake. Indeed, the refusal to find state action in *Logiodice* appears motivated in part by the Court’s conclusion that the private school’s failure to hold a hearing on the student’s suspension—which the majority describes as one of “the small arguable unfairnesses that are part of life”—should not constitute a due process violation.

4. Private Prisons

This leaves application of the proposed analysis to privately operated prisons. Currently, private prisons are a fairly straightforward case for finding state action on public function grounds. Notwithstanding the lengthy historical pedigree of private involvement in incarceration, today punishment and the legitimate use of physical coercion are seen as exclusive state prerogatives. Similarly, under the proposed analysis, the use of private prisons easily constitutes a delegation of government power, whether because of the agency relationship created by contracts between governments and prison operators or because incarceration is inherently an exercise of government power.

The more interesting issue is whether alternative mechanisms exist that would adequately preserve constitutional accountability without direct application of constitutional constraints to private prison operators. The most obvious candidate is an administrative complaint system under which inmates would have extensive ability to challenge private prisons’ decisions and actions. But the powers wielded by prison operators and guards are too pervasive and the interactions between prison staff and inmates too inherently discretionary to be adequately policed solely by a system of indirect controls. This conclusion draws support from the Court’s decision in *Malesko*. There, in refusing to
imply a private right of action for damages for constitutional violations against the 
private prison operator, the Court underscored the availability of alternative means for 
enforcing constitutional requirements—direct injunctive relief and actions against 
individual prison guards—in addition to the availability of administrative review.

An alternative mechanism would be regulatory or contractual provisions requiring private 
prisons to respect the constitutional rights of inmates and authorizing inmates to bring 
suit against private prisons to challenge alleged violations. Here, direct protections 
against private abuse are provided. However, this route often would preclude suits against 
private prisons in federal court, as unless the prisons were operating on the behalf of the 
federal government, actions based on these provisions would be rooted solely in state 
law. Absent a basis in federal statutory law, ensuring access to federal courts in this 
context requires finding private prisons to be state actors directly subject to constitutional 
constraints. Given the historical importance of federal courts in improving conditions in 
public prisons, the prudential case for direct application of constitutional constraints is 
particularly strong.

Thus, privately operated prisons represent one of the instances where under the proposed 
analysis private entities might still be deemed state actors. Yet the proposed analysis 
could lead to different results in regard to other forms of prison privatization. For 
example, in West, the Court found that a private doctor providing medical care to inmates 
pursuant to a contract with a public prison was a state actor. This contractual arrangement 
clearly qualifies as a private delegation that created an agency relationship. Moreover, on 
the facts of West, that delegation was not adequately structured to preserve constitutional 
accountability: Inmates could obtain medical care only through the doctor selected by the 
prison, and the government did not provide an administrative or statutory means for 
challenging the doctor’s decisions. However, under the proposed analysis, the result 
would be invalidation of the delegation rather than authorization of a constitutional suit 
against the private doctor, with the government being liable for any inadequate care the 
doctor had provided. The proposed analysis and current state action doctrine would have 
diverged even further if North Carolina had instead subsidized medical treatment by a 
doctor of a prisoner’s choosing and allowed prisoners a meaningful choice among 
providers. In that case, no doctor would exercise a monopoly over access to medical care 
and the individual prisoner’s ability to choose providers would suffice to address 
constitutional accountability concerns.

CONCLUSION

The vision of clearly distinct public and private realms embodied in current state action 
doctrine bears little resemblance to the modern administrative state, with its endless and 
varied reliance upon private actors to perform the tasks of government. The result of this 
disjuncture is a profound challenge to the principle of constitutional accountability. 
Current doctrine targets government involvement, either involvement in specific private 
acts or pervasive entwinement with the private entity more generally, for constitutional 
condemnation. But such government involvement is rarely present, and also is not the 
real constitutional concern. Far more threatening is the potential that private actors wield 
broad discretion over government programs with insufficient government oversight and
also (as a result) outside of constitutional controls. Reforming existing doctrine to reflect current administrative reality is critical, particularly given recent moves to increase public-private partnerships and expand the authority delegated to private entities. At the same time, however, simply targeting existing doctrine as underinclusive is too facile, for it fails to take account of why contemporary constitutional law is so accepting of private delegations, notwithstanding the obvious constitutional dangers. A major reason for that phenomenon is the perception that gains in constitutional accountability come at the expense of government regulatory freedom.

This Article argues not only that current state action doctrine is singularly deficient, but further that it is possible to secure constitutional accountability without destroying regulatory flexibility. The key to so doing lies in accepting private exercises of government power as a reality of modern government and then replacing state action’s focus on whether constitutional requirements should apply to private actors. The real constitutional concern should be instead whether private delegations of government power are adequately structured to preserve constitutional accountability. No reason exists to hold private entities directly subject to constitutional constraints when alternative mechanisms exist by which individuals can enforce constitutional limits on government power. Moreover, direct constitutional scrutiny of private entities is also generally inappropriate when such mechanisms are lacking. The fundamental constitutional flaw in this context lies in the nature of the delegation involved—the government’s decision to delegate government power to private entities without ensuring that constitutional controls will be preserved. Responsibility to rectify such an unconstitutional delegation should lie with the political branches, not the courts.
<table>
<thead>
<tr>
<th>Case</th>
<th>Current State Action Doctrine</th>
<th>Proposed Private Delegation Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blum v. Yaretsky</td>
<td>No state action</td>
<td>Unconstitutional private delegation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘The delegation to nursing homes of power to determine extent of Medicaid beneficiaries’ need for nursing services creates an agency relationship, and this delegation is not adequately structured to preserve constitutional accountability: No mechanism exists by which individual beneficiaries can challenge nursing homes’ care determinations; in addition, homes exercise quasi-monopoly powers over residents.</td>
</tr>
<tr>
<td>Rendell-Baker v. Kohn</td>
<td>No state action</td>
<td>Constitutional private delegation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The delegation to private school of responsibility to provide education to special needs students creates an agency relationship, but, at least with regard to teacher terminations, informal state oversight is adequate to preserve constitutional accountability.</td>
</tr>
<tr>
<td>West v. Atkins</td>
<td>State action</td>
<td>Unconstitutional private delegation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The delegation to a private doctor to provide medical services to inmates creates an agency relationship, and this delegation is not adequately structured to preserve constitutional accountability: The government undertook no oversight of the doctor’s medical decisions; private law remedies such as malpractice suits do not support a claim for failure to provide medical services; and the doctor exercised a monopoly over inmate access to medical care.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘The statutory authorization for private medical insurers to decline to pay workers’ compensation claims absent a finding of medical necessity does not create an agency relationship.</td>
</tr>
<tr>
<td>Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass’n</td>
<td>State action</td>
<td>Unconstitutional private delegation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘The prior state delegation of exclusive control over interscholastic athletics, combined with public school officials’ de facto control of the Association, creates an agency relationship, and this delegation is not adequately structured to preserve constitutional accountability: The Association’s decisions were not subject to government review; no alternative means of enforcing constitutional requirements existed; and the Association exercised substantial authority over high school athletics in the state.</td>
</tr>
<tr>
<td>San Francisco Arts &amp; Athletics, Inc. v. United States Olympic Comm.</td>
<td>No state action</td>
<td>Unconstitutional private delegation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whether Congress’ delegation to the USOC of exclusive oversight over amateur sports created an agency relationship with the federal government is a close question. This delegation, as evidenced by Congress’ chartering of the USOC and granting it direct access to federal funds, served federal policy goals; however, the government bore no statutory responsibility to provide oversight of amateur sports or of the United States’ representation at the Olympics. Regardless, the grant of exclusive oversight represents a clear delegation of government power, because the USOC is authorized to regulate amateur athletics and take enforcement action against individual athletes and organizations. While procedural requirements imposed on the USOC by statute may be adequate to meet due process requirements, mechanisms to enforce other constitutional limits appear lacking.</td>
</tr>
</tbody>
</table>
§ 5. Why Are There Firms?

§ 5.1. ADAM SMITH, THE WEALTH OF NATIONS (1776), Book I, Chapter I. Of the Division of Labour

The greatest improvements in the productive powers of labour, and the greater part of the skill, dexterity, and judgment, with which it is anywhere directed, or applied, seem to have been the effects of the division of labour. . . .

To take an example, therefore, from a very trifling manufacture, but one in which the division of labour has been very often taken notice of, the trade of a pin-maker: a workman not educated to this business (which the division of labour has rendered a distinct trade), nor acquainted with the use of the machinery employed in it (to the invention of which the same division of labour has probably given occasion), could scarce, perhaps, with his utmost industry, make one pin in a day, and certainly could not make twenty. But in the way in which this business is now carried on, not only the whole work is a peculiar trade, but it is divided into a number of branches, of which the greater part are likewise peculiar trades. One man draws out the wire; another straightens it; a third cuts it; a fourth points it; a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on is a peculiar business; to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufactories, are all performed by distinct hands, though in others the same man will sometimes perform two or three of them.

I have seen a small manufactory of this kind, where ten men only were employed, and where some of them consequently performed two or three distinct operations. But though they were very poor, and therefore but indifferently accommodated with the necessary machinery, they could, when they exerted themselves, make among them about twelve pounds of pins in a day. There are in a pound upwards of four thousand pins of a middling size. Those ten persons, therefore, could make among them upwards of forty-eight thousand pins in a day. Each person, therefore, making a tenth part of forty-eight thousand pins, might be considered as making four thousand eight hundred pins in a day. . . .

In every other art and manufacture, the effects of the division of labour are similar to what they are in this very trifling one, though, in many of them, the labour can neither be so much subdivided, nor reduced to so great a simplicity of operation. The division of labour, however, so far as it can be introduced, occasions, in every art, a proportionable increase of the productive powers of labour. The separation of different trades and employments from one another, seems to have taken place in consequence of this advantage. This separation, too, is generally carried furthest in those countries which enjoy the highest degree of industry and improvement; what is the work of one man, in a rude state of society, being generally that of several in an improved one. In every improved society, the farmer is generally nothing but a farmer; the manufacturer, nothing
but a manufacturer. The labour, too, which is necessary to produce any one complete
manufacture, is almost always divided among a great number of hands. How many
different trades are employed in each branch of the linen and woollen manufactures, from
the growers of the flax and the wool, to the bleachers and smoothers of the linen, or to the
dyers and dressers of the cloth!

The nature of agriculture, indeed, does not admit of so many subdivisions of labour, nor
of so complete a separation of one business from another, as manufactures. It is
impossible to separate so entirely the business of the grazier from that of the corn-farmer,
as the trade of the carpenter is commonly separated from that of the smith. The spinner is
almost always a distinct person from the weaver; but the ploughman, the harrower, the
sower of the seed, and the reaper of the corn, are often the same. The occasions for those
different sorts of labour returning with the different seasons of the year, it is impossible
that one man should be constantly employed in any one of them. This impossibility of
making so complete and entire a separation of all the different branches of labour
employed in agriculture, is perhaps the reason why the improvement of the productive
powers of labour, in this art, does not always keep pace with their improvement in
manufactures. The most opulent nations, indeed, generally excel all their neighbours in
agriculture as well as in manufactures; but they are commonly more distinguished by
their superioritity in the latter than in the former. Their lands are in general better
cultivated, and having more labour and expense bestowed upon them, produce more in
proportion to the extent and natural fertility of the ground. But this superioritity of produc-
is seldom much more than in proportion to the superioritity of labour and expense. In
agriculture, the labour of the rich country is not always much more productive than that
of the poor; or, at least, it is never so much more productive, as it commonly is in
manufactures. The corn of the rich country, therefore, will not always, in the same degree
of goodness, come cheaper to market than that of the poor. . . .

This great increase in the quantity of work, which, in consequence of the division of
labour, the same number of people are capable of performing, is owing to three different
circumstances; first, to the increase of dexterity in every particular workman; secondly, to
the saving of the time which is commonly lost in passing from one species of work to
another; and, lastly, to the invention of a great number of machines which facilitate and
abridge labour, and enable one man to do the work of many.

First, the improvement of the dexterity of the workmen, necessarily increases the quantity
of the work he can perform; and the division of labour, by reducing every man’s business
to some one simple operation, and by making this operation the sole employment of his
life, necessarily increases very much the dexterity of the workman. A common smith,
who, though accustomed to handle the hammer, has never been used to make nails, if,
upon some particular occasion, he is obliged to attempt it, will scarce, I am assured, be
able to make above two or three hundred nails in a day, and those, too, very bad ones. A
smith who has been accustomed to make nails, but whose sole or principal business has
not been that of a nailer, can seldom, with his utmost diligence, make more than eight
hundred or a thousand nails in a day. I have seen several boys, under twenty years of age,
who had never exercised any other trade but that of making nails, and who, when they
exerted themselves, could make, each of them, upwards of two thousand three hundred nails in a day.

The making of a nail, however, is by no means one of the simplest operations. The same person blows the bellows, stirs or mends the fire as there is occasion, heats the iron, and forges every part of the nail: in forging the head, too, he is obliged to change his tools. The different operations into which the making of a pin, or of a metal button, is subdivided, are all of them much more simple, and the dexterity of the person, of whose life it has been the sole business to perform them, is usually much greater. The rapidity with which some of the operations of those manufactures are performed, exceeds what the human hand could, by those who had never seen them, be supposed capable of acquiring.

Secondly, The advantage which is gained by saving the time commonly lost in passing from one sort of work to another, is much greater than we should at first view be apt to imagine it. It is impossible to pass very quickly from one kind of work to another, that is carried on in a different place, and with quite different tools. A country weaver, who cultivates a small farm, must loose a good deal of time in passing from his loom to the field, and from the field to his loom. When the two trades can be carried on in the same workhouse, the loss of time is, no doubt, much less. It is, even in this case, however, very considerable. A man commonly saunters a little in turning his hand from one sort of employment to another. When he first begins the new work, he is seldom very keen and hearty; his mind, as they say, does not go to it, and for some time he rather trifles than applies to good purpose. The habit of sauntering, and of indolent careless application, which is naturally, or rather necessarily, acquired by every country workman who is obliged to change his work and his tools every half hour, and to apply his hand in twenty different ways almost every day of his life, renders him almost always slothful and lazy, and incapable of any vigorous application, even on the most pressing occasions. Independent, therefore, of his deficiency in point of dexterity, this cause alone must always reduce considerably the quantity of work which he is capable of performing.

Thirdly, and lastly, everybody must be sensible how much labour is facilitated and abridged by the application of proper machinery. It is unnecessary to give any example. I shall only observe, therefore, that the invention of all those machines by which labour is to much facilitated and abridged, seems to have been originally owing to the division of labour. Men are much more likely to discover easier and readier methods of attaining any object when the whole attention of their minds is directed towards that single object, than when it is dissipated among a great variety of things. But, in consequence of the division of labour, the whole of every man’s attention comes naturally to be directed towards some one very simple object. It is naturally to be expected, therefore, that some one or other of those who are employed in each particular branch of labour should soon find out easier and readier methods of performing their own particular work, whenever the nature of it admits of such improvement.

A great part of the machines made use of in those manufactures in which labour is most subdivided, were originally the invention of common workmen, who, being each of them employed in some very simple operation, naturally turned their thoughts towards finding
out easier and readier methods of performing it. Whoever has been much accustomed to visit such manufactures, must frequently have been shewn very pretty machines, which were the inventions of such workmen, in order to facilitate and quicken their own particular part of the work. In the first fire engines [i.e., steam engines], a boy was constantly employed to open and shut alternately the communication between the boiler and the cylinder, according as the piston either ascended or descended. One of those boys, who loved to play with his companions, observed that, by tying a string from the handle of the valve which opened this communication to another part of the machine, the valve would open and shut without his assistance, and leave him at liberty to divert himself with his play-fellows. One of the greatest improvements that has been made upon this machine, since it was first invented, was in this manner the discovery of a boy who wanted to save his own labour.

All the improvements in machinery, however, have by no means been the inventions of those who had occasion to use the machines. Many improvements have been made by the ingenuity of the makers of the machines, when to make them became the business of a peculiar trade; and some by that of those who are called philosophers, or men of speculation, whose trade it is not to do any thing, but to observe every thing, and who, upon that account, are often capable of combining together the powers of the most distant and dissimilar objects. In the progress of society, philosophy or speculation becomes, like every other employment, the principal or sole trade and occupation of a particular class of citizens. Like every other employment, too, it is subdivided into a great number of different branches, each of which affords occupation to a peculiar tribe or class of philosophers; and this subdivision of employment in philosophy, as well as in every other business, improve dexterity, and saves time. Each individual becomes more expert in his own peculiar branch, more work is done upon the whole, and the quantity of science is considerably increased by it.

It is the great multiplication of the productions of all the different arts, in consequence of the division of labour, which occasions, in a well-governed society, that universal opulence which extends itself to the lowest ranks of the people. Every workman has a great quantity of his own work to dispose of beyond what he himself has occasion for; and every other workman being exactly in the same situation, he is enabled to exchange a great quantity of his own goods for a great quantity or, what comes to the same thing, for the price of a great quantity of theirs. He supplies them abundantly with what they have occasion for, and they accommodate him as amply with what he has occasion for, and a general plenty diffuses itself through all the different ranks of the society.

Observe the accommodation of the most common artificer or day-labourer in a civilized and thriving country, and you will perceive that the number of people, of whose industry a part, though but a small part, has been employed in procuring him this accommodation, exceeds all computation. The woollen coat, for example, which covers the day-labourer, as coarse and rough as it may appear, is the produce of the joint labour of a great multitude of workmen. The shepherd, the sorter of the wool, the wool-comber or carder, the dyer, the scribbler, the spinner, the weaver, the fuller, the dresser, with many others, must all join their different arts in order to complete even this homely production. How many merchants and carriers, besides, must have been employed in transporting the
materials from some of those workmen to others who often live in a very distant part of
the country? How much commerce and navigation in particular, how many ship-builders,
sailors, sail-makers, rope-makers, must have been employed in order to bring together the
different drugs made use of by the dyer, which often come from the remotest corners of
the world? What a variety of labour, too, is necessary in order to produce the tools of the
meanest of those workmen! To say nothing of such complicated machines as the ship of
the sailor, the mill of the fuller, or even the loom of the weaver, let us consider only what
a variety of labour is requisite in order to form that very simple machine, the shears with
which the shepherd clips the wool. The miner, the builder of the furnace for smelting the
ore, the seller of the timber, the burner of the charcoal to be made use of in the smelting-
house, the brickmaker, the bricklayer, the workmen who attend the furnace, the
millwright, the forger, the smith, must all of them join their different arts in order to
produce them.

Were we to examine, in the same manner, all the different parts of his dress and
household furniture, the coarse linen shirt which he wears next his skin, the shoes which
cover his feet, the bed which he lies on, and all the different parts which compose it, the
kitchen-grate at which he prepares his victuals, the coals which he makes use of for that
purpose, dug from the bowels of the earth, and brought to him, perhaps, by a long sea and
a long land-carriage, all the other utensils of his kitchen, all the furniture of his table, the
knives and forks, the earthen or pewter plates upon which he serves up and divides his
victuals, the different hands employed in preparing his bread and his beer, the glass
window which lets in the heat and the light, and keeps out the wind and the rain, with all
the knowledge and art requisite for preparing that beautiful and happy invention, without
which these northern parts of the world could scarce have afforded a very comfortable
habitation, together with the tools of all the different workmen employed in producing
those different conveniences; if we examine, I say, all these things, and consider what a
variety of labour is employed about each of them, we shall be sensible that, without the
assistance and co-operation of many thousands, the very meanest person in a civilized
country could not be provided, even according to what we very falsely imagine the easy
and simple manner in which he is commonly accommodated. Compared, indeed, with the
more extravagant luxury of the great, his accommodation must no doubt appear
extremely simple and easy; and yet it may be true, perhaps, that the accommodation of an
European prince does not always so much exceed that of an industrious and frugal
peasant, as the accommodation of the latter exceeds that of many an African king, the
absolute master of the lives and liberties of ten thousand naked savages.

---

§ 5.2. R.H. Coase, The Nature of the Firm, 4 ECONOMICA (n.s.) 386 (1937)

Economic theory has suffered in the past from a failure to state clearly its assumptions.
Economists in building up a theory have often omitted to examine the foundations on
which it was erected. This examination is, however, essential not only to prevent the
misunderstanding and needless controversy which arise from a lack of knowledge of the

275
Draft — do not distribute!
assumptions on which a theory is based, but also because of the extreme importance for
economics of good judgment in choosing between rival sets of assumptions. For instance,
it is suggested that the use of the word “firm” in economics may be different from the use
of the term by the “plain man.” Since there is apparently a trend in economic theory
towards starting analysis with the individual firm and not with the industry, it is all the
more necessary not only that a clear definition of the word “firm” should be given, but
that its difference from a firm in the “real world,” if it exists, should be made clear. . . .

I

It is convenient if, in searching for a definition of a firm, we first consider the economic
system as it is normally treated by the economist. Let us consider the description of the
For its current operation it is under no central control, it needs no central survey. Over the
whole range of human activity and human need, supply is adjusted to demand, and
production to consumption, by a process that is automatic, elastic and responsive.” An
economist thinks of the economic system as being coordinated by the price mechanism,
and society becomes not an organisation but an organism. The economic system “works
itself.” This does not mean that there is no planning by individuals. These exercise
foresight and choose between alternatives. This is necessarily so if there is to be order in
the system. But this theory assumes that the direction of resources is dependent directly
on the price mechanism. Indeed, it is often considered to be an objection to economic
planning that it merely tries to do what is already done by the price mechanism.

Sir Arthur Salter’s description, however, gives a very incomplete picture of our economic
system. Within a firm, the description does not fit at all. . . . If a workman moves from
department Y to department X, he does not go because of a change in relative prices, but
because he is ordered to do so. Those who object to economic planning on the grounds
that the problem is solved by price movements can be answered by pointing out that there
is planning within our economic system which is quite different from the individual
planning mentioned above and which is akin to what is normally called economic
planning. The example given above is typical of a large sphere in our modern economic
system. Of course, this fact has not been ignored by economists. . . . But in view of the
fact that it is usually argued that co-ordination will be done by the price mechanism, why
is such organisation necessary? . . . Outside the firm, price movements direct production,
which is coordinated through a series of exchange transactions on the market. Within a
firm these market transactions are eliminated, and in place of the complicated market
structure with exchange transactions is substituted the entrepreneur–co-ordinator, who
directs production. It is clear that these are alternative methods of co-ordinating
production. Yet, having regard to the fact that, if production is regulated by price
movements, production could be carried on without any organisation at all, well might
we ask, why is there any organisation? . . .

It can, I think, be assumed that the distinguishing mark of the firm is the supersession of
the price mechanism. . . .
In view of the fact that, while economists treat the price mechanism as a coordinating instrument, they also admit the coordinating function of the “entrepreneur,” it is surely important to enquire why co-ordination is the work of the price mechanism in one case and of the entrepreneur in another. The purpose of this paper is to bridge what appears to be a gap in economic theory between the assumption (made for some purposes) that resources are allocated by means of the price mechanism and the assumption (made for other purposes) that this allocation is dependent on the entrepreneur–co-ordinator. We have to explain the basis on which, in practice, this choice between alternatives is effected.

II

Our task is to attempt to discover why a firm emerges at all in a specialized exchange economy. The price mechanism (considered purely from the side of the direction of resources) might be superseded if the relationship which replaced it was desired for its own sake. This would be the case, for example, if some people preferred to work under the direction of some other person. Such individuals would accept less in order to work under someone, and firms would arise naturally from this. But it would appear that this cannot be a very important reason, for it would rather seem that the opposite tendency is operating if one judges from the stress normally laid on the advantage of “being one’s own master.” Of course, if the desire was not to be controlled but to control, to exercise power over others, then people might be willing to give up something in order to direct others; that is, they would be willing to pay others more than they could get under the price mechanism in order to be able to direct them. But this implies that those who direct pay in order to be able to do this and are not paid to direct, which is clearly not true in the majority of cases. Firms might also exist if purchasers preferred commodities which are produced by firms to those not so produced; but even in spheres where one would expect such preferences (if they exist) to be of negligible importance, firms are to be found in the real world. Therefore there must be other elements involved.

The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism. The most obvious cost of “organising” production through the price mechanism is that of discovering what the relevant prices are. This cost may be reduced but it will not be eliminated by the emergence of specialists who will sell this information. The costs of negotiating and concluding a separate contract for each exchange transaction which takes place on a market must also be taken into account. Again, in certain markets, e.g., produce exchanges, a technique is devised for minimizing these contract costs; but they are not eliminated. It is true that contracts are not eliminated when there is a firm, but they are greatly reduced. A factor of production (or the owner thereof) does not have to make a series of contracts with the factors with whom he is cooperating within the firm, as would be necessary, of course, if this co-operation were a direct result of the working of the price mechanism. For this series of contracts is substituted one. At this stage, it is important to note the character of the contract into which a factor enters that is employed within a firm. The contract is one whereby the factor, for a certain remuneration (which may be fixed or fluctuating), agrees to obey the directions of an entrepreneur within certain limits. The essence of the contract is that it
should only state the limits to the powers of the entrepreneur. Within these limits, he can therefore direct the other factors of production.

There are, however, other disadvantages—or costs—of using the price mechanism. It may be desired to make a long-term contract for the supply of some article or service. This may be due to the fact that if one contract is made for a longer period, instead of several shorter ones, then certain costs of making each contract will be avoided. Or, owing to the risk attitude of the people concerned, they may prefer to make a long rather than a short-term contract. Now, owing to the difficulty of forecasting, the longer the period of the contract is for the supply of the commodity or service, the less possible, and indeed, the less desirable it is for the person purchasing to specify what the other contracting party is expected to do. It may well be a matter of indifference to the person supplying the service or commodity which of several courses of action is taken, but not the purchaser of that service or commodity. But the purchaser will not know which of these several courses he will want the supplier to take. Therefore, the service which is being provided is expressed in general terms, the exact details being left until a later date. All that is stated in the contract is the limits to what the person supplying the commodity or service is expected to do. The details of what the supplier is expected to do is not stated in the contract but is decided later by the purchaser. When the direction of resources (within the limits of the contract) becomes dependent on the buyer in this way, that relationship which I term a “firm” may be obtained. A firm is likely, therefore, to emerge in those cases where a very short-term contract would be unsatisfactory. It is obviously of more importance in the case of services—labour—than it is in the case of the buying of commodities. In the case of commodities, the main items can be stated in advance and the details which will be decided later will be of minor significance.

We may sum up this section of the argument by saying that the operation of a market costs something and that, by forming an organisation and allowing some authority (an “entrepreneur”) to direct the resources, certain marketing costs are saved. The entrepreneur has to carry out his function at less cost, taking into account the fact, that he may get factors of production at a lower price than the market transactions which he supersedes, because it is always possible to revert to the open market if he fails to do this.

The question of uncertainty is one which is often considered to be very relevant to the study of the equilibrium of the firm. It seems improbable that a firm would emerge without the existence of uncertainty ... .

These, then, are the reasons why organisations such as firms exist in a specialized exchange economy in which it is generally assumed that the distribution of resources is “organised” by the price mechanism. A firm, therefore, consists of the system of relationships which comes into existence when the direction of resources is dependent on an entrepreneur.

The approach which has just been sketched would appear to offer an advantage, in that it is possible to give a scientific meaning to what is meant by saying that a firm gets larger or smaller. A firm becomes larger as additional transactions (which could be exchange transactions coordinated through the price mechanism) are organised by the entrepreneur,
and it becomes smaller as he abandons the organisation of such transactions. The question which arises is whether it is possible to study the forces which determine the size of the firm. Why does the entrepreneur not organise one less transaction or one more? . . .

It was suggested that the introduction of the firm was due primarily to the existence of marketing costs. A pertinent question to ask would appear to be . . . why, if by organising one can eliminate certain costs and in fact reduce the cost of production, are there any market transactions at all? Why is not all production carried on by one big firm? There would appear to be certain possible explanations.

First, as a firm gets larger there may be decreasing returns to the entrepreneur function, that is, the costs of organising additional transactions within the firm may rise. Naturally, a point must be reached where the costs of organising an extra transaction within the firm are equal to the costs involved in carrying out the transaction in the open market or to the costs of organising by another entrepreneur. Second, it may be that, as the transactions which are organised increase, the entrepreneur fails to place the factors of production in the uses where their value is greatest, that is, fails to make the best use of the factors of production. Again, a point must be reached where the loss through the waste of resources is equal to the marketing costs of the exchange transaction in the open market or to the loss if the transaction was organised by another entrepreneur. Finally, the supply price of one or more of the factors of production may rise, because the “other advantages” of a small firm are greater than those of a large firm. Of course, the actual point where the expansion of the firm ceases might be determined by a combination of the factors mentioned above. The first two reasons given most probably correspond to the economists’ phrase of “diminishing returns to management.” . . .

Other things being equal, therefore, a firm will tend to be larger:

(a) the less the costs of organising and the slower these costs rise with an increase in the transactions organised;

(b) the less likely the entrepreneur is to make mistakes and the smaller the increase in mistakes with an increase in the transactions organised;

(c) the greater the lowering (or the less the rise) in the supply price of factors of production to firms of larger size.

Apart from variations in the supply price of factors of production to firms of different sizes, it would appear that the costs of organising and the losses through mistakes will increase with an increase in the spatial distribution of the transactions organised, in the dissimilarity of the transactions, and in the probability of changes in the relevant prices. As more transactions are organised by an entrepreneur, it would appear that the transactions would tend to be either different in kind or different in place. This furnishes an additional reason why efficiency will tend to decrease as the firm gets larger. Inventions which tend to bring factors of production nearer together, by lessening spatial distribution, tend to increase the size of the firm. Changes like the telephone and the
telegraph, which tend to reduce the cost of organising spatially, will tend to increase the size of the firm. All changes which improve managerial technique will tend to increase the size of the firm . . . .

V

. . . [T]he definition we have given is one which closely approximates the firm as it is considered in the real world.

Our definition is, therefore, realistic. Is it manageable? This ought to be clear. When we are considering how large a firm will be, the principle of marginalism works smoothly. The question always is, will it pay to bring an extra exchange transaction under the organising authority? At the margin, the costs of organising within the firm will be equal either to the costs of organising in another firm or to the costs involved in leaving the transaction to be “organised” by the price mechanism. Business men will be constantly experimenting, controlling more or less, and in this way equilibrium will be maintained. This gives the position of equilibrium for static analysis. But it is clear that the dynamic factors are also of considerable importance, and an investigation of the effect changes have on the cost of organising within the firm and on marketing costs generally will enable one to explain why firms get larger and smaller. We thus have a theory of moving equilibrium.

The above analysis would also appear to have clarified the relationship between initiative and enterprise and management. Initiative means forecasting and operates through the price mechanism by the making of new contracts. Management proper merely reacts to price changes, rearranging the factors of production under its control. That the business man normally combines both functions is an obvious result of the marketing costs which were discussed above. Finally, this analysis enables us to state more exactly what is meant by the “marginal product” of the entrepreneur. But an elaboration of this point would take us far from our comparatively simple task of definition and clarification.

§ 5.3. Armen A. Alchian & Harold Demsetz, Production, Information Costs, and Economic Organization, 62 AM. ECON. REV. 777 (1972)

The mark of a capitalistic society is that resources are owned and allocated by such nongovernmental organizations as firms, households, and markets. Resource owners increase productivity through cooperative specialization and this leads to the demand for economic organizations which facilitate cooperation. When a lumber mill employs a cabinetmaker, cooperation between specialists is achieved within a firm, and when a cabinetmaker purchases wood from a lumberman, the cooperation takes place across markets (or between firms). Two important problems face a theory of economic organization—to explain the conditions that determine whether the gains from
specialization and cooperative production can better be obtained within an organization like the firm, or across markets, and to explain the structure of the organization.

It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can “punish” you only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. That is exactly all that any employer can do. He can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products. What then is the content of the presumed power to manage and assign workers to various tasks? Exactly the same as one little consumer’s power to manage and assign his grocer to various tasks. The single consumer can assign his grocer to the task of obtaining whatever the customer can induce the grocer to provide at a price acceptable to both parties. That is precisely all that an employer can do to an employee. To speak of managing, directing, or assigning workers to various tasks is a deceptive way of noting that the employer continually is involved in renegotiation of contracts on terms that must be acceptable to both parties. Telling an employee to type this letter rather than to file that document is like my telling a grocer to sell me this brand of tuna rather than that brand of bread. I have no contract to continue to purchase from the grocer and neither the employer nor the employee is bound by any contractual obligations to continue their relationship. Long-term contracts between employer and employee are not the essence of the organization we call a firm. My grocer can count on my returning day after day and purchasing his services and goods even with the prices not always marked on the goods—because I know what they are—and he adapts his activity to conform to my directions to him as to what I want each day . . . he is not my employee.

Wherein then is the relationship between a grocer and his employee different from that between a grocer and his customers? It is in a team use of inputs and a centralized position of some party in the contractual arrangements of all other inputs. It is the centralized contractual agent in a team productive process—not some superior authoritarian directive or disciplinary power. Exactly what is a team process and why does it induce the contractual form, called the firm? These problems motivate the inquiry of this paper.

I. THE METERING PROBLEM

The economic organization through which input owners cooperate will make better use of their comparative advantages to the extent that it facilitates the payment of rewards in accord with productivity. If rewards were random, and without regard to productive effort, no incentive to productive effort would be provided by the organization; and if rewards were negatively correlated with productivity the organization would be subject to sabotage. Two key demands are placed on an economic organization—metering input productivity and metering rewards.
Metering problems sometimes can be resolved well through the exchange of products across competitive markets, because in many situations markets yield a high correlation between rewards and productivity. If a farmer increases his output of wheat by 10 percent at the prevailing market price, his receipts also increase by 10 percent. This method of organizing economic activity meters the output directly, reveals the marginal product and apportions the rewards to resource owners in accord with that direct measurement of their outputs. The success of this decentralized, market exchange in promoting productive specialization requires that changes in market rewards fall on those responsible for changes in output.

The classic relationship in economics that runs from marginal productivity to the distribution of income implicitly assumes the existence of an organization, be it the market or the firm, that allocates rewards to resources in accord with their productivity. The problem of economic organization, the economical means of metering productivity and rewards, is not confronted directly in the classical analysis of production and distribution. Instead, that analysis tends to assume sufficiently economic—or zero cost—means, as if productivity automatically created its reward. We conjecture the direction of causation is the reverse—the specific system of rewarding which is relied upon stimulates a particular productivity response. If the economic organization meters poorly, with rewards and productivity only loosely correlated, then productivity will be smaller; but if the economic organization meters well productivity will be greater. What makes metering difficult and hence induces means of economizing on metering costs?

II. TEAM PRODUCTION

Two men jointly lift heavy cargo into trucks. Solely by observing the total weight loaded per day, it is impossible to determine each person’s marginal productivity. With team production it is difficult, solely by observing total output, to either define or determine each individual’s contribution to this output of the cooperating inputs. The output is yielded by a team, by definition, and it is not a sum of separable outputs of each of its members.

Usual explanations of the gains from cooperative behavior rely on exchange and production in accord with the comparative advantage specialization principle with separable additive production. However, as suggested above there is a source of gain from cooperative activity involving working as a team, wherein individual cooperating inputs do not yield identifiable, separate products which can be summed to measure the total output. For this cooperative productive activity, here called “team” production, measuring marginal productivity and making payments in accord therewith is more expensive by an order of magnitude than for separable production functions.

Team production, to repeat, is production in which

1) several types of resources are used and

2) the product is not a sum of separable outputs of each cooperating resource.

An additional factor creates a team organization problem—
3) not all resources used in team production belong to one person.

We do not inquire into why all the jointly used resources are not owned by one person, but instead into the types of organization, contracts, and informational and payment procedures used among owners of teamed inputs. With respect to the one-owner case, perhaps it is sufficient merely to note that (a) slavery is prohibited, (b) one might assume risk aversion as a reason for one person’s not borrowing enough to purchase all the assets or sources of services rather than renting them, and (c) the purchase-resale spread may be so large that costs of short-term ownership exceed rental costs. Our problem is viewed basically as one of organization among different people, not of the physical goods or services, however much there must be selection and choice of combination of the latter.

How can the members of a team be rewarded and induced to work efficiently? In team production, marginal products of cooperative team members are not so directly and separably (i.e., cheaply) observable. What a team offers to the market can be taken as the marginal product of the team but not of the team members. The costs of metering or ascertaining the marginal products of the team’s members is what calls forth new organizations and procedures. Clues to each input’s productivity can be secured by observing behavior of individual inputs. When lifting cargo into the truck, how rapidly does a man move to the next piece to be loaded, how many cigarette breaks does he take, does the item being lifted tilt downward toward his side?

If detecting such behavior were costless, neither party would have an incentive to shirk, because neither could impose the cost of his shirking on the other (if their cooperation was agreed to voluntarily). But since costs must be incurred to monitor each other, each input owner will have more incentive to shirk when he works as part of a team, than if his performance could be monitored easily or if he did not work as a team. If there is a net increase in productivity available by team production, net of the metering cost associated with disciplining the team, then team production will be relied upon rather than a multitude of bilateral exchange of separable individual outputs.

Both leisure and higher income enter a person’s utility function. Hence, each person should adjust his work and realized reward so as to equate the marginal rate of substitution between leisure and production of real output to his marginal rate of substitution in consumption. That is, he would adjust his rate of work to bring his demand prices of leisure and output to equality with their true costs. However, with detection, policing, monitoring, measuring or metering costs, each person will be induced to take more leisure, because the effect of relaxing on his realized (reward) rate of substitution between output and leisure will be less than the effect on the true rate of substitution. His realized cost of leisure will fall more than the true cost of leisure, so he “buys” more leisure (i.e., more nonpecuniary reward).

If his relaxation cannot be detected perfectly at zero cost, part of its effects will be borne by others in the team, thus making his realized cost of relaxation less than the true total cost to the team. The difficulty of detecting such actions permits the private costs of his actions to be less than their full costs. Since each person responds to his private realizable rate of substitution (in production) rather than the true total (i.e., social) rate, and so long
as there are costs for other people to detect his shift toward relaxation, it will not pay (them) to force him to readjust completely by making him realize the true cost. Only enough efforts will be made to equate the marginal gains of detection activity with the marginal costs of detection; and that implies a lower rate of productive effort and more shirking than in a costless monitoring, or measuring, world.

In a university, the faculty use office telephones, paper, and mail for personal uses beyond strict university productivity. The university administrators could stop such practices by identifying the responsible person in each case, but they can do so only at higher costs than administrators are willing to incur. The extra costs of identifying each party (rather than merely identifying the presence of such activity) would exceed the savings from diminished faculty “turpitudinal peccadillos.” So the faculty is allowed some degree of “privileges, perquisites, or fringe benefits.” And the total of the pecuniary wages paid is lower because of this irreducible (at acceptable costs) degree of amenity-seizing activity. Pay is lower in pecuniary terms and higher in leisure, conveniences, and ease of work. But still every person would prefer to see detection made more effective (if it were somehow possible to monitor costlessly) so that he, as part of the now more effectively producing team, could thereby realize a higher pecuniary pay and less leisure. If everyone could, at zero cost, have his reward-realized rate brought to the true production possibility real rate, all could achieve a more preferred position. But detection of the responsible parties is costly; that cost acts like a tax on work rewards. Viable shirking is the result.

What forms of organizing team production will lower the cost of detecting “performance” (i.e., marginal productivity) and bring personally realized rates of substitution closer to true rates of substitution? Market competition, in principle, could monitor some team production. (It already organizes teams.) Input owners who are not team members can offer, in return for a smaller share of the team’s rewards, to replace excessively (i.e., overpaid) shirking members. Market competition among potential team members would determine team membership and individual rewards. There would be no team leader, manager, organizer, owner, or employer. For such decentralized organizational control to work, outsiders, possibly after observing each team’s total output, can speculate about their capabilities as team members and, by a market competitive process, revised teams with greater productive ability will be formed and sustained. Incumbent members will be constrained by threats of replacement by outsiders offering services for lower reward shares or offering greater rewards to the other members of the team. Any team member who shirked in the expectation that the reduced output effect would not be attributed to him will be displaced if his activity is detected. Teams of productive inputs, like business units, would evolve in apparent spontaneity in the market—without any central organizing agent, team manager, or boss.

But completely effective control cannot be expected from individualized market competition for two reasons. First, for this competition to be completely effective, new challengers for team membership must know where, and to what extent, shirking is a serious problem, i.e., know they can increase net output as compared with the inputs they replace. To the extent that this is true it is probably possible for existing fellow team members to recognize the shirking. But, by definition, the detection of shirking by
observing team output is costly for team production. Secondly, assume the presence of
detection costs, and assume that in order to secure a place on the team a new input owner
must accept a smaller share of rewards (or a promise to produce more). Then his
incentive to shirk would still be at least as great as the incentives of the inputs replaced,
because he still bears less than the entire reduction in team output for which he is
responsible.

III. THE CLASSICAL FIRM

One method of reducing shirking is for someone to specialize as a monitor to check the
input performance of team members. But who will monitor the monitor? One constraint
on the monitor is the aforesaid market competition offered by other monitors, but for
reasons already given, that is not perfectly effective. Another constraint can be imposed
on the monitor: give him title to the net earnings of the team, net of payments to other
inputs. If owners of cooperating inputs agree with the monitor that he is to receive any
residual product above prescribed amounts (hopefully, the marginal value products of the
other inputs), the monitor will have an added incentive not to shirk as a monitor.

Specialization in monitoring plus reliance on a residual claimant status will reduce
shirking; but additional links are needed to forge the firm of classical economic theory.

How will the residual claimant monitor the other inputs?

We use the term monitor to connote several activities in addition to its disciplinary
connotation. It connotes measuring output performance, apportioning rewards, observing
the input behavior of inputs as means of detecting or estimating their marginal
productivity and giving assignments or instructions in what to do and how to do it. (It
also includes, as we shall show later, authority to terminate or revise contracts.) Perhaps
the contrast between a football coach and team captain is helpful. The coach selects
strategies and tactics and sends in instructions about what plays to utilize. The captain is
essentially an observer and reporter of the performance at close hand of the members.
The latter is an inspector-steward and the former a supervisor manager. For the present
all these activities are included in the rubric “monitoring.” All these tasks are, in
principle, negotiable across markets, but we are presuming that such market measurement
of marginal productivities and job reassignments are not so cheaply performed for team
production. And in particular our analysis suggests that it is not so much the costs of
spontaneously negotiating contracts in the markets among groups for team production as
it is the detection of the performance of individual members of the team that calls for the
organization noted here.

The specialist who receives the residual rewards will be the monitor of the members of
the team (i.e., will manage the use of cooperative inputs). The monitor earns his residual
through the reduction in shirking that he brings about, not only by the prices that he
agrees to pay the owners of the inputs, but also by observing and directing the actions or
uses of these inputs. Managing or examining the ways to which inputs are used in team
production is a method of metering the marginal productivity of individual inputs to the
team’s output.
To discipline team members and reduce shirking, the residual claimant must have power to revise the contract terms and incentives of individual members without having to terminate or alter every other input’s contract. Hence, team members who seek to increase their productivity will assign to the monitor not only the residual claimant right but also the right to alter individual membership and performance on the team. Each team member, of course, can terminate his own membership (i.e., quit the team), but only the monitor may unilaterally terminate the membership of any of the other members without necessarily terminating the team itself or his association with the team; and he alone can expand or reduce membership, alter the mix of membership, or sell the right to be the residual claimant-monitor of the team. It is this entire bundle of rights: 1) to be a residual claimant; 2) to observe input behavior; 3) to be the central party common to all contracts with inputs; 4) to alter the membership of the team; and 5) to sell these rights, that defines the ownership (or the employer) of the classical (capitalist, free-enterprise) firm. The coalescing of these rights has arisen, our analysis asserts, because it resolves the shirking-information problem of team production better than does the noncentralized contractual arrangement.

The relationship of each team member to the owner of the firm (i.e., the party common to all input contracts and the residual claimant) is simply a “quid pro quo” contract. Each makes a purchase and sale. The employee “orders” the owner of the team to pay him money in the same sense that the employer directs the team member to perform certain acts. The employee can terminate the contract as readily as can the employer, and longterm contracts, therefore, are not an essential attribute of the firm. Nor are “authoritarian,” “dictational,” or “fiat” attributes relevant to the conception of the firm or its efficiency.

In summary, two necessary conditions exist for the emergence of the firm on the prior assumption that more than pecuniary wealth enter utility functions:

1) It is possible to increase productivity through team-oriented production, a production technique for which it is costly to directly measure the marginal outputs of the cooperating inputs. This makes it more difficult to restrict shirking through simple market exchange between cooperating inputs.

2) It is economical to estimate marginal productivity by observing or specifying input behavior.

The simultaneous occurrence of both these preconditions leads to the contractual organization of inputs, known as the classical capitalist firms with (a) joint input production, (b) several input owners, (c) one party who is common to all the contracts of the joint inputs, (d) who has rights to renegotiate any input’s contract independently of contracts with other input owners, (e) who holds the residual claim, and (f) who has the right to sell his central contractual residual status.
Other Theories of the Firm

At this juncture, as an aside, we briefly place this theory of the firm in the contexts of [that] offered by Ronald Coase. . . . Our view of the firm is not necessarily inconsistent with Coase’s; we attempt to go further and identify refutable implications. Coase’s penetrating insight is to make more of the fact that markets do not operate costlessly, and he relies on the cost of using markets to form contracts as his basic explanation for the existence of firms. We do not disagree with the proposition that . . . the higher is the cost of transacting across markets the greater will be the comparative advantage of organizing resources within the firm; it is a difficult proposition to disagree with or to refute. We could with equal ease subscribe to a theory of the firm based on the cost of managing, for surely it is true that . . . the lower is the cost of managing the greater will be the comparative advantage of organizing resources within the firm. To move the theory forward, it is necessary to know what is meant by a firm and to explain the circumstances under which the cost of “managing” resources is low relative to the cost of allocating resources through market transaction. The conception of and rationale for the classical firm that we propose takes a step down the path pointed out by Coase toward that goal. Consideration of team production, team organization, difficulty in metering outputs, and the problem of shirking are important to our explanation but, so far as we can ascertain, not in Coase’s. Coase’s analysis insofar as it had heretofore been developed would suggest open-ended contracts but does not appear to imply anything more—neither the residual claimant status nor the distinction between employee and subcontractor status (nor any of the implications indicated below). And it is not true that employees are generally employed on the basis of long-term contractual arrangements any more than on a series of short-term or indefinite length contracts.

The importance of our proposed additional elements is revealed, for example, by the explanation of why the person to whom the control monitor is responsible receives the residual, and also by our later discussion of the implications about the corporation, partnerships, and profit sharing. These alternative forms for organization of the firm are difficult to resolve on the basis of market transaction costs only. Our exposition also suggests a definition of the classical firm—something crucial that was heretofore absent.

In addition, sometimes a technological development will lower the cost of market transactions while, at the same time, it expands the role of the firm. When the “putting out” system was used for weaving, inputs were organized largely through market negotiations. With the development of efficient central sources of power, it became economical to perform weaving in proximity to the power source and to engage in team production. The bringing in of weavers surely must have resulted in a reduction in the cost of negotiating (forming) contracts. Yet, what we observe is the beginning of the factory system in which inputs are organized within a firm. Why? The weavers did not simply move to a common source of power that they could tap like an electric line, purchasing power while they used their own equipment. Now team production in the joint use of equipment became more important. The measurement of marginal productivity, which now involved interactions between workers, especially through their joint use of machines, became more difficult though contract negotiating cost was reduced, while managing the behavior of inputs became easier because of the increased centralization of
activity. The firm as an organization expanded even though the cost of transactions was
reduced by the advent of centralized power. The same could be said for modern assembly
lines. Hence the emergence of central power sources expanded the scope of productive
activity in which the firm enjoyed a comparative advantage as an organizational form.

Although we have emphasized team production as creating a costly metering task and
have treated team production as an essential (necessary?) condition for the firm, would
not other obstacles to cheap metering also call forth the same kind of contractual
arrangement here denoted as a firm? For example, suppose a farmer produces wheat in an
easily ascertained quantity but with subtle and difficult to detect quality variations
determined by how the farmer grew the wheat. A vertical integration could allow a
purchaser to control the farmer’s behavior in order to more economically estimate
productivity. But this is not a case of joint or team production, unless “information” can
be considered part of the product. (While a good case could be made for that broader
conception of production, we shall ignore it here.) Instead of forming a firm, a buyer can
contract to have his inspector on the site of production, just as home builders contract
with architects to supervise building contracts; that arrangement is not a firm. Still, a firm
might be organized in the production of many products wherein no team production or
jointness of use of separately owned resources is involved.

This possibility rather clearly indicates a broader, or complementary, approach to that
which we have chosen.

1) As we do in this paper, it can be argued that the firm is the particular policing
device utilized when joint team production is present. If other sources of high
policing costs arise, as in the wheat case just indicated, some other form of
contractual arrangement will be used. Thus to each source of informational cost
there may be a different type of policing and contractual arrangement.

2) On the other hand, one can say that where policing is difficult across markets,
various forms of contractual arrangements are devised, but there is no reason for
that known as the firm to be uniquely related or even highly correlated with team
production, as defined here. It might be used equally probably and viably for
other sources of high policing cost. We have not intensively analyzed other
sources, and we can only note that our current and readily revisable conjecture is
that 1) is valid, and has motivated us in our current endeavor.

In any event, the test of the theory advanced here is to see whether the conditions we
have identified are necessary for firms to have long-run viability rather than merely births
with high infant mortality. Conglomerate firms or collections of separate production
agencies into one owning organization can be interpreted as an investment trust or
investment diversification device . . . . A holding company can be called a firm, because
of the common association of the word firm with any ownership unit that owns income
sources. The term firm as commonly used is so turgid of meaning that we can not hope to
explain every entity to which the name is attached in common or even technical
literature. Instead, we seek to identify and explain a particular contractual arrangement
induced by the cost of information factors analyzed in this paper.

288
Draft — do not distribute!
V. TEAM SPIRIT AND LOYALTY

Every team member would prefer a team in which no one, not even himself, shirked. Then the true marginal costs and values could be equated to achieve more preferred positions. If one could enhance a common interest in nonshirking in the guise of a team loyalty or team spirit, the team would be more efficient. In those sports where team activity is most clearly exemplified, the sense of loyalty and team spirit is most strongly urged. Obviously the team is better, with team spirit and loyalty, because of the reduced shirking—not because of some other feature inherent in loyalty or spirit as such.

Corporations and business firms try to instill a spirit of loyalty. This should not be viewed simply as a device to increase profits by over-working or misleading the employees, nor as an adolescent urge for belonging. It promotes a closer approximation to the employees’ potentially available true rates of substitution between production and leisure and enables each team member to achieve a more preferred situation. The difficulty, of course, is to create economically that team spirit and loyalty. It can be preached with an aura of moral code of conduct—a morality with literally the same basis as the ten commandments—to restrict our conduct toward what we would choose if we bore our full costs.

VI. KINDS OF INPUTS OWNED BY THE FIRM

To this point the discussion has examined why firms, as we have defined them, exist? That is, why is there an owner-employer who is the common party to contracts with other owners of inputs in team activity? The answer to that question should also indicate the kind of the jointly used resources likely to be owned by the central-owner-monitor and the kind likely to be hired from people who are not team-owners. Can we identify characteristics or features of various inputs that lead to their being hired or to their being owned by the firm? . . .

[The hire-or-own decision can be explained by, among other things, the factors] developed above to explain the existence of the institution known as the firm—the costs of detecting output performance. When a durable resource is used it will have a marginal product and a depreciation. Its use requires payment to cover at least use-induced depreciation; unless that user cost is specifically detectable, payment for it will be demanded in accord with expected depreciation. And we can ascertain circumstances for each. An indestructible hammer with a readily detectable marginal product has zero user cost. But suppose the hammer were destructible and that careless (which is easier than careful) use is more abusive and causes greater depreciation of the hammer. Suppose in addition the abuse is easier to detect by observing the way it is used than by observing only the hammer after its use, or by measuring the output scored from a hammer by a laborer. If the hammer were rented and used in the absence of the owner, the depreciation would be greater than if the use were observed by the owner and the user charged in accord with the imposed depreciation. (Careless use is more likely than careful use—if one does not pay for the greater depreciation.) An absentee owner would therefore ask for a higher rental price because of the higher expected user cost than if the item were used by the owner. The expectation is higher because of the greater difficulty of observing
specific user cost, by inspection of the hammer after use. Renting is therefore in this case more costly than owner use. This is the valid content of the misleading expressions about ownership being more economical than renting—ignoring all other factors that may work in the opposite direction, like tax provision, short-term occupancy and capital risk avoidance.

Better examples are tools of the trade. Watch repairers, engineers, and carpenters tend to own their own tools especially if they are portable. Trucks are more likely to be employee owned rather than other equally expensive team inputs because it is relatively cheap for the driver to police the care taken in using a truck. Policing the use of trucks by a nondriver owner is more likely to occur for trucks that are not specialized to one driver, like public transit busses.

The factor with which we are concerned here is one related to the costs of monitoring not only the gross product performance of an input but also the abuse or depreciation inflicted on the input in the course of its use. If depreciation or user cost is more cheaply detected when the owner can see its use than by only seeing the input before and after, there is a force toward owner use rather than renting. Resources whose user cost is harder to detect when used by someone else, tend on this count to be owner-used. Absentee ownership, in the lay language, will be less likely. Assume momentarily that labor service cannot be performed in the absence of its owner. The labor owner can more cheaply monitor any abuse of himself than if somehow labor-services could be provided without the labor owner observing its mode of use or knowing what was happening. Also his incentive to abuse himself is increased if he does not own himself.

The similarity between the preceding analysis and the question of absentee landlordism and of sharecropping arrangements is no accident. The same factors which explain the contractual arrangements known as a firm help to explain the incidence of tenancy, labor hiring or sharecropping.

VII. FIRMS AS A SPECIALIZED MARKET INSTITUTION FOR COLLECTING, COLLATING, AND SELLING INPUT INFORMATION

The firm serves as a highly specialized surrogate market. Any person contemplating a joint-input activity must search and detect the qualities of available joint inputs. He could contact an employment agency, but that agency in a small town would have little advantage over a large firm with many inputs. The employer, by virtue of monitoring many inputs, acquires special superior information about their productive talents. This aids his directive (i.e., market hiring) efficiency. He “sells” his information to employee-inputs as he aids them in ascertaining good input combinations for team activity. Those who work as employees or who rent services to him are using him to discern superior combinations of inputs. Not only does the director-employer “decide” what each input will produce, he also estimates which heterogeneous inputs will work together jointly more efficiently, and he does this in the context of a privately owned market for forming teams. The department store is a firm and is a superior private market. People who shop and work in one town can as well shop and work in a privately owned firm.
This marketing function is obscured in the theoretical literature by the assumption of homogeneous factors. Or it is tacitly left for individuals to do themselves via personal market search, much as if a person had to search without benefit of specialist retailers. Whether or not the firm arose because of this efficient information service, it gives the director-employer more knowledge about the productive talents of the team’s inputs, and a basis for superior decisions about efficient or profitable combinations of those heterogeneous resources.

In other words, opportunities for profitable team production by inputs already within the firm may be ascertained more economically and accurately than for resources outside the firm. Superior combinations of inputs can be more economically identified and formed from resources already used in the organization than by obtaining new resources (and knowledge of them) from the outside. Promotion and revision of employee assignments (contracts) will be preferred by a firm to the hiring of new inputs. To the extent that this occurs there is reason to expect the firm to be able to operate as a conglomerate rather than persist in producing a single product. Efficient production with heterogeneous resources is a result not of having better resources but in knowing more accurately the relative productive performances of those resources. Poorer resources can be paid less in accord with their inferiority; greater accuracy of knowledge of the potential and actual productive actions of inputs rather than having high productivity resources makes a firm (or an assignment of inputs) profitable.

VIII. SUMMARY

While ordinary contracts facilitate efficient specialization according to comparative advantage, a special class of contracts among a group of joint inputs to a team production process is commonly used for team production. Instead of multilateral contracts among all the joint inputs’ owners, a central common party to a set of bilateral contracts facilitates efficient organization of the joint inputs in team production. The terms of the contracts form the basis of the entity called the firm—especially appropriate for organizing team production processes.

Team productive activity is that in which a union, or joint use, of inputs yields a larger output than the sum of the products of the separately used inputs. This team production requires—like all other production processes—an assessment of marginal productivities if efficient production is to be achieved. Nonseparability of the products of several differently owned joint inputs raises the cost of assessing the marginal productivities of those resources or services of each input owner. Monitoring or metering the productivities to match marginal productivities to costs of inputs and thereby to reduce shirking can be achieved more economically (than by across market bilateral negotiations among inputs) in a firm.

The essence of the classical firm is identified here as a contractual structure with:

1) joint input production;

2) several input owners;
3) one party who is common to all the contracts of the joint inputs;

4) who has rights to renegotiate any input’s contract independently of contracts with other input owners;

5) who holds the residual claim; and

6) who has the right to sell his central contractual residual status.

The central agent is called the firm’s owner and the employer. No authoritarian control is involved; the arrangement is simply a contractual structure subject to continuous renegotiation with the central agent. The contractual structure arises as a means of enhancing efficient organization of team production. In particular, the ability to detect shirking among owners of jointly used inputs in team production is enhanced (detection costs are reduced) by this arrangement and the discipline (by revision of contracts) of input owners is made more economic.

Testable implications are suggested by the analysis of different types of organizations—nonprofit, proprietary for profit, unions, cooperatives, partnerships, and by the kinds of inputs that tend to be owned by the firm in contrast to those employed by the firm. We conclude with a highly conjectural but possibly significant interpretation. As a consequence of the flow of information to the central party (employer), the firm takes on the characteristic of an efficient market in that information about the productive characteristics of a large set of specific inputs is now more cheaply available. Better recombinations or new uses of resources can be more efficiently ascertained than by the conventional search through the general market. In this sense inputs compete with each other within and via a firm rather than solely across markets as conventionally conceived. Emphasis on interfirm competition obscures intrafirm competition among inputs. Conceiving competition as the revelation and exchange of knowledge or information about qualities, potential uses of different inputs in different potential applications indicates that the firm is a device for enhancing competition among sets of input resources as well as a device for more efficiently rewarding the inputs. In contrast to markets and cities which can be viewed as publicly or nonowned market places, the firm can be considered a privately owned market; if so, we could consider the firm and the ordinary market as competing types of markets, competition between private proprietary markets and public or communal markets. Could it be that the market suffers from the defects of communal property rights in organizing and influencing uses of valuable resources?

§ 5.4. Oliver E. Williamson, *The Vertical Integration of Production: Market Failure Considerations*, AM. ECON. REV., May 1971 (special issue), at 112

The study of vertical integration has presented difficulties at both theoretical and policy levels of analysis. That vertical integration has never enjoyed a secure place in value
theory is attributable to the fact that, under conventional assumptions, it is an anomaly: if the costs of operating competitive markets are zero, “as is usually assumed in our theoretical analysis” (Arrow, 1969), why integrate?

Policy interest in vertical integration has been concerned mainly with the possibility that integration can be used strategically to achieve anticompetitive effects. In the absence of a more substantial theoretical foundation, vertical integration, as a public policy matter, is typically regarded as having dubious if not outright antisocial properties. Technological interdependencies or, possibly, observational economies, constitute the principal exceptions.

The technological interdependency argument is both the most familiar and the most straight-forward: successive processes which, naturally, follow immediately in time and place dictate certain efficient manufacturing configurations; these, in turn, are believed to have common ownership implications. Such technical complementarity is probably more important in flow process operations (chemicals, metals, etc.) than in separable component manufacture. The standard example is the integration of iron and steel-making, where thermal economies are said to be available through integration. It is commonly held that where “integration does not have this physical or technical aspect—as it does not, for example, in integrating the production of assorted components with the assembly of those components—the case for cost savings from integration is generally much less clear” (Bain, 1968).

There is, nevertheless, a distinct unease over the argument. This is attributable, probably, to a suspicion that the firm is more than a simple efficiency instrument, in the usual scale economies and least-cost factor proportions senses of the term, but also possesses coordinating potential that sometimes transcends that of the market. It is the burden of the present argument that this suspicion is warranted. In more numerous respects than are commonly appreciated, the substitution of internal organization for market exchange is attractive less on account of technological economies associated with production but because of what may be referred to broadly as “transactional failures” in the operation of markets for intermediate goods. This substitution of internal organization for market exchange will be referred to as “internalization.”

The two principal prior contributions on which the argument relies are Coase’s seminal discussion on “The Nature of The Firm” (1937) and Arrow’s more recent review of market versus nonmarket allocation (1969). As will be evident, I agree with Malmgren (1961) that the analysis of transaction costs is uninteresting under fully stationary conditions and that only when the need to make unprogrammed adaptations is introduced does the market versus internal organization issue become engaging.

But while Malmgren finds that the advantage of the firm inures in its capacity to control information and achieve plan consistency among interdependent activities, which may be regarded as an information processing advantage, I mainly emphasize the differential incentive and control properties of firms in relation to markets. This is not to suggest that information processing considerations are unimportant, but rather that these incompletely
characterize the distinctive properties of firms that favor internal organization as a market substitute.

I. INTERNAL ORGANIZATION: AFFIRMATIVE ASPECTS

A complete treatment of vertical integration requires that the limits as well as the powers of internal organization be assessed. As the frictions associated with administrative coordination become progressively more severe, recourse to market exchange becomes more attractive . . . . It is beyond the scope of this paper, however, to examine the organizational failure aspect of the vertical integration question. Rather it is simply asserted that, mainly on account of bounded rationality and greater confidence in the objectivity of market exchange in comparison with bureaucratic processes, market intermediation is generally to be preferred over internal supply in circumstances in which markets may be said to “work well.”

The properties of the firm that commend internal organization as a market substitute would appear to fall into three categories: incentives, controls, and what may be referred to broadly as “inherent structural advantages.” In an incentive sense, internal organization attenuates the aggressive advocacy that epitomizes arms length bargaining. Interests, if not perfectly harmonized, are at least free of representations of a narrowly opportunistic sort; in any viable group, of which the firm is one, the range of admissible intraorganizational behavior is bounded by considerations of alienation. In circumstances, therefore, where protracted bargaining between independent parties to a transaction can reasonably be anticipated, internalization becomes attractive.

Perhaps the most distinctive advantage of the firm, however, is the wider variety and greater sensitivity of control instruments that are available for enforcing intrafirm in comparison with interfirm activities (Williamson, 1970). Not only does the firm have the constitutional authority and low-cost access to the requisite data which permit it to perform more precise own-performance evaluations (of both a contemporaneous and ex post variety) than can a buyer, but its reward and penalty instruments (which include selective use of employment, promotion, remuneration, and internal resource allocation processes) are more refined.

Especially relevant in this connection is that, when conflicts develop, the firm possesses a comparatively efficient conflict resolution machinery. To illustrate, fiat is frequently a more efficient way to settle minor conflicts (say differences of interpretation) than is haggling or litigation. Interorganizational conflict can be settled by fiat only rarely, if at all. For one thing, it would require the parties to agree on an impartial arbitrator, which agreement itself may be costly to secure. It would also require that rules of evidence and procedure be established. If, moreover, the occasion for such interorganizational settlements were to be common, the form of organization converges in effect to vertical integration, with the arbiter becoming a manager in fact if not in name. By contrast, intraorganizational settlements by fiat are common (Whinston, 1964).

The firm may also resort to internalization on account of economies of information exchange. Some of these may be due to structural differences between firms and markets.
Others, however, reduce ultimately to incentive and control differences between internal and market organization. It is widely accepted, for example, that communication with respect to complex matters is facilitated by a common training and experience and if a compact code has developed in the process. Repeated interpersonal interactions may permit even further economies of communication; subtle nuances may come through in familiar circumstances which in an unfamiliar relationship could be achieved only with great effort. Still, the drawing of an organizational boundary need not, by itself, prevent intensely familiar relations from developing between organizations. Put differently, but for the goal and control differences described above, the informational advantages of internal over market organization are not, in this respect, apparent. Claims of informational economies thus should distinguish between economies that are attributable to information flows per se (structure) and those which obtain on account of differential veracity effects (see Part D, Section 2).

II. MARKET FAILURE CONSIDERATIONS

What are referred to here as market failures are failures only in the limited sense that they involve transaction costs that can be attenuated by substituting internal organization for market exchange. The argument proceeds in five stages. The first three are concerned with characterizing a successively more complex bargaining environment in which small numbers relations obtain. The last two deal with the special structural advantages which, either naturally or because of prevailing institutional rules, the firm enjoys in relation to the market.

A. Static Markets

Consider an industry that produces a multicomponent product, assume that some of these components are specialized (industry specific), and assume further that among these there are components for which the economies of scale in production are large in relation to the market. The market, then, will support only a few efficient sized producers for certain components.

A monopolistic excess of price over cost under market procurement is commonly anticipated in these circumstances—although, as Demsetz (1968) has noted, this need not obtain if there are large numbers of suppliers willing and able to bid at the initial contract award stage. Assume, however, that large numbers bidding is not feasible. The postulated conditions then afford an “apparent” incentive for assemblers to integrate backward or suppliers to integrate forward. Two different cases can be distinguished: bilateral monopoly (oligopoly) and competitive assembly with monopolistic supply. The former is considered here; the latter is treated in Part C.

Bilateral monopoly requires that both price and quantity be negotiated. Both parties stand to benefit, naturally, by operating on rather than off the contract curve—which here corresponds to the joint profit maximizing quantity (Fellner, 1947). But this merely establishes the amount to be exchanged. The terms at which this quantity will be traded still need to be determined. Any price consistent with nonnegative profits to both parties is feasible. Bargaining can be expected to ensue. Haggling will presumably continue until
the marginal private net benefits are perceived by one of the parties to be zero. Although this haggling is jointly (and socially) unproductive, it constitutes a source of private pecuniary gain. Being, nevertheless, a joint profit drain, an incentive to avoid these costs, if somehow this could be arranged, is set up.

One possible adaptation is to internalize the transaction through vertical integration; but a once-for-all contract might also be negotiated. In a perfectly static environment (one that is free of disturbances of all kinds), these may be regarded with indifference: the former involves settlement on component supply price while merger requires agreement on asset valuation. Bargaining skills will presumably be equally important in each instance (indeed, a component price can be interpreted in asset valuation terms and conversely). Thus, although vertical integration may occur under these conditions, there is nothing in the nature of the problem that requires such an outcome.

A similar argument in these circumstances also applies to adaptation against externalities: joint profit considerations dictate that the affected parties reach an accommodation, but integration holds no advantage over once-for-all contracts in a perfectly static environment.

Transforming the relationship from one of bilateral monopoly to one of bilateral oligopoly broadens the range of bargaining alternatives, but the case for negotiating a merger agreement in relation to a once-for-all contract is not differentially affected on this account. The static characterization of the problem, apparently, will have to be relaxed if a different result is to be reached.

**B. Contractual Incompleteness**

Let the above conditions be enriched to include the stipulation that the product in question is technically complex and that periodic redesign and/or volume changes are made in response to changing environmental conditions. Also relax the assumption that large numbers bidding at the initial contract award stage is infeasible. Three alternative supply arrangements can be considered: a once-for-all contract, a series of short-term contracts, and vertical integration.

The dilemma posed by once-for-all contracts is this: lest independent parties interpret contractual ambiguities to their own advantage, which differences can be resolved only by haggling or, ultimately, litigation, contingent supply relations ought exhaustively to be stipulated. But exhaustive stipulation, assuming that it is feasible, is itself costly. Thus although, if production functions were known, appropriate responses to final demand or factor price changes might be deduced, the very costliness of specifying the functions and securing agreement discourages the effort. The problem is made even more severe where a changing technology poses product redesign issues. Here it is doubtful that, despite great effort and expense, contractual efforts reasonably to comprehend the range of possible outcomes will be successful. An adaptive, sequential decision process is thus indicated. If, however, contractual revisions or amendments are regarded as an occasion to bargain opportunistically, which predictably they will be, the purchaser will defer and accumulate adaptations, if by packaging them in complex combinations their true value
can better be disguised; some adaptations may be foregone altogether. The optimal sequential decision-making process can in these respects be distorted.

Short-term contracts, which would facilitate adaptive, sequential decision-making, might therefore be preferred. These pose problems, however, if either (1) efficient supply requires investment in special-purpose, long-life equipment, or (2) the winner of the original contract acquires a cost advantage, say by reason of “first mover” advantages (such as unique location or learning, including the acquisition of undisclosed or proprietary technical and managerial procedures and task-specific labor skills).

The problem with condition (1) is that optimal investment considerations favor the award of a long-term contract so as to permit the supplier confidently to amortize his investment. But, as indicated, long term contracts pose adaptive, sequential decision-making problems. Thus optimal investment and optimal sequential adaptation processes are in conflict in this instance.

It might be argued that condition (2) poses no problems since initial bidders will fully reflect in their original bids all relevant factors. Thus, although anticipated downstream cost advantages (where downstream is used both here and subsequently in the sense of time rather than place) will give rise to small numbers competition for downstream supply, competition at the initial award stage is sufficient to assure that only competitive returns will be realized over the entire supply interval. One might expect, therefore, that the low bidder would come in at a price below cost in the first period, set price at the level of alternative supply price in later periods, and earn normal returns over-all. Appropriate changes can be introduced easily at the recontracting interval.

A number of potential problems are posed, however. For one thing, unless the total supply requirements are stipulated, “buying in” strategies are risky. Also, and related, the alternative supply price is not independent of the terms that the buyer may subsequently offer to rivals. Moreover, alternative supply price is merely an upper bound; an aggressive buyer may attempt to obtain a price at the level of current costs on each round. Haggling could be expected to ensue. Short-term contracts thus experience what may be serious limitations in circumstances where nontrivial first-mover advantages obtain.

In consideration, therefore, of the problems that both long and short-term contracts are subject to, vertical integration may well be indicated. The conflict between efficient investment and efficient sequential decision-making is thereby avoided. Sequential adaptations become an occasion for cooperative adjustment rather than opportunistic bargaining; risks may be attenuated; differences between successive stages can be resolved more easily by the internal control machinery.

It is relevant to note that the technological interdependency condition involving flow process economies between otherwise separable stages of production is really a special case of the contractual incompleteness argument. The contractual dilemma is this: On the one hand, it may be prohibitively costly, if not infeasible, to specify contractually the full range of contingencies and stipulate appropriate responses between stages. On the other hand, if the contract is seriously incomplete in these respects but, once the original
negotiations are settled, the contracting parties are locked into a bilateral exchange, the divergent interests between the parties will predictably lead to individually opportunistic behavior and joint losses. The advantages of integration thus are not that technological (flow process) economies are unavailable to nonintegrated firms, but that integration harmonizes interests (or reconciles differences, often by fiat) and permits an efficient (adaptive, sequential) decision process to be utilized. More generally, arguments favorable to integration that turn on “supply reliability” considerations commonly reduce to the contractual incompleteness issue.

C. Strategic Misrepresentation Risk

Contractual incompleteness problems develop where there is ex ante but not necessarily ex post uncertainty. Strategic misrepresentation risks are serious where there is uncertainty in both respects. Not only is the future uncertain but it may not be possible, except at great cost, for an outside agency to establish accurately what has transpired after the fact. The advantages of internalization reside in the facts that the firm’s ex post access to the relevant data is superior, it attenuates the incentives to exploit uncertainty opportunistically, and the control machinery that the firm is able to activate is more selective.

1. Affirmative Occasions for Integration

Three affirmative occasions to integrate on account of strategic misrepresentation risk and two potentially anticompetitive consequences of integration can be identified.

a. Moral Hazard

The problem here arises because of the conjoining of inharmonious incentives with uncertainty—or, as Arrow puts it (1969), it is due to the “confounding of risks and decisions.” To illustrate, consider the problem of contracting for an item the final cost and/or performance of which is subject to uncertainty. One possibility is for the supplier to bear the uncertainty. But, he will undertake a fixed price contract to deliver a specified result the costs of which are highly uncertain only after attaching a risk premium to the price. Assume that the buyer regards this premium as excessive and is prepared on this account to bear the risk himself. The risk can easily be shifted by offering a cost-plus contract. But this impairs the incentives of the supplier to achieve least-cost performance; the supplier may reallocate his assets in such a way as to favor other work to the disadvantage of the cost-plus contract.

Thus, although, if commitments were self-enforcing, it might often be institutionally most efficient to divide the functions of risk bearing and contract execution (that is, cost-plus contracts would have ideal properties), specialization is discouraged by interest disparities. At a minimum, the buyer may insist on monitoring the supplier’s work. In contrast therefore to a fixed-price contract, where it is sufficient to evaluate end-product performance, cost-plus contracts, because they expose the buyer to risks of inefficient (high cost) contract execution, require that both inputs and outputs be evaluated.
Internalization does not eliminate the need for input evaluation. Rather, the advantage of internalization, for input monitoring purposes, resides in the differential ease with which controls are exercised. An external agency, by design, lacks recourse to the internal control machinery: proposed remedies require the consent of the contractor and then are highly circumscribed; unrestricted access by the buyer to the contractor’s internal control machinery (including selective use of employment, promotion, remuneration, and internal resource allocation processes) is apt to be denied. In consideration of the costs and limitations of input monitoring by outsiders, the buyer may choose instead to bear the risk and perform the work himself. The buyer thus internalizes, through backward vertical integration, a transaction which, but for uncertainty, would move through the market. A cost-type contract for internal procurement is arranged.

*b. Externalities/Imputation*

The externality issue can be examined in two parts. First, has a secure, unambiguous, and “appropriate” assignment of property rights been made? Second, are the accounting costs of imputing costs and benefits substantial? If answers to these questions are affirmative and negative respectively, appropriability problems will not become an occasion for vertical integration. Where these conditions are not satisfied, however, integration may be indicated.

The assignment aspect of this matter is considered in Part E below. Here it is assumed that an efficacious assignment of property rights has been made and that only the expense of imputing costs and benefits is at issue. But indeed this is apt often to be the more serious problem. High imputation expenses which discourage accurate metering introduce ambiguity into transactions. Did party A affect party B and if so in what degree? In the absence of objective, low cost standards, opposed interests can be expected to evaluate these effects differently. Internalization, which permits protracted (and costly) disputes over these issues to be avoided, may on this account be indicated.

c. *Variable Proportions Distortions*

Consider the case where the assembly stage will support large numbers; fewness appears only in component supply. Whether monopolistic supply prices provide an occasion for vertical integration in these circumstances depends both on production technology and policing expense. Variable proportions at the assembly stage afford opportunities for nonintegrated assemblers to adapt against monopolistically priced components by substituting competitively priced factors (McKenzie, 1951). Although conceivably the monopolistic component supplier could stipulate, as a condition of sale, that fixed proportions in assembly should prevail, the effectiveness of such stipulations is to be questioned—since, ordinarily, the implied enforcement costs will be great. Where substitution occurs, inefficient factor proportions, with consequent welfare losses, will result. The private (and social) incentives to integrate so as to reduce total costs by restoring efficient factor combinations are evident.
2. Anticompetitive Consequences

Anticompetitive effects of two types are commonly attributed to integration: price discrimination and barriers to entry (cf. Stigler, 1968).

a. Price Discrimination

The problem here is first to discover differential demand elasticities, and secondly to arrange for sale in such a way as to preclude reselling. Users with highly elastic demands which purchase the item at a low price must not be able to service inelastic demand customers by acting as a middleman; all sales must be final. Although vertical integration may facilitate the discovery of differential elasticities, it is mainly with respect to the non-resale condition that it is regarded as especially efficacious.

Integration, nevertheless, is a relatively extreme response. Moreover, price discrimination is clearly practiced in some commodities without recourse to vertical integration (witness electricity and telephone service). What are the distinguishing factors? Legality considerations aside, presumably it is the cost of enforcing (policing) terms of the contract that are at issue. Some commodities apparently have self-enforcing properties—which may obtain on account of high storage and repacking costs or because reselling can not be arranged inconspicuously. The absence of self-enforcing (policing) properties is what makes vertical integration attractive as a means of accomplishing discrimination.

b. Entry Barrier Effects

That the vertical integration of production might be used effectively to bar entry is widely disputed. Bork (1969) argues that “In general, if greater than competitive profits are to be made in an industry, entry should occur whether the entrant has to come in at both levels at once or not. I know of no theory of imperfections in the capital market which would lead suppliers of capital to avoid areas of higher return to seek areas of lower return.” But the issue is not one of profit avoidance but rather involves cost incidence. If borrowers are confronted by increasingly adverse rates as they increase their finance requirements, which Hirshleifer suggests is a distinct possibility (1970), cost may not be independent of vertical structure.

Assuming that vertical integration has the effect of increasing capital requirements, the critical issues are to what extent and for what reasons the supply curve of finance behaves in the way postulated. The following conjecture is offered as a partial explanation: unable to monitor the performance of large, complex organizations in any but the crudest way or to effect management displacement easily except on evidence of seriously discreditable error, investors demand larger returns as finance requirements become progressively greater, ceteris paribus. Thus the costs of policing against the contingency that managers will operate a rival enterprise opportunistically are, on this argument, at least partly responsible for the reputed behavior of the supply curve of capital. In consideration of this state of affairs, established firms may use vertical integration strategically to increase finance requirements and thereby to discourage entry if potential entrants feel compelled,
as a condition of successful entry, to adopt the prevailing structure—as they may if the industry is highly concentrated.

D. Information Processing Effects

As indicated in Section I, one of the advantages of the firm is that it realizes economies of information exchange. These may manifest themselves as information impactedness, observational economies, or what Malmgren (1961) refers to as the “convergence of expectations.”

1. Information Impactedness

Richardson illustrates the problems of information impactedness by reference to an entrepreneur who was willing to offer long-term contracts (at normal rates of return, presumably) but which contracts others were unprepared to accept because they were not convinced that he had “the ability, as well as the will, to fulfill them. He may have information sufficient to convince himself that this is the case, but others may not” (Richardson, 1960). He goes on to observe that the perceived risks of the two parties may be such as to make it difficult to negotiate a contract that offers commensurate returns to each; objective risks are augmented by contractual risks in these circumstances. Integration undertaken for this reason is akin to self-insurance by individuals who know themselves to be good risks but are priced out of the insurance market because of their inability, at low cost, to “reveal” this condition to insurers.

2. Observational Economies

As Radner indicates, “the acquisition of information often involves a ‘set-up cost’; i.e., the resources needed to obtain the information may be independent of the scale of the production process in which the information is used” (Radner, 1970). Although Radner apparently had horizontal firm size implications in mind, the argument also has relevance for vertical integration. If a single set of observations can be made that is of relevance to a related series of production stages, vertical integration may be efficient.

Still, the question might be raised, why common ownership? Why not an independent observational agency that sells information to all comers? Or, if the needed information is highly specialized, why not a joint venture? Alternatively, what inhibits efficient information exchange between successive stages of production according to contract? In relation, certainly, to the range of intermediate options potentially available, common ownership appears to be an extreme response. What are the factors which favor this outcome?

One of the problems with contracts is that of specifying terms. But even if terms could be reached, there is still a problem of policing the agreement. To illustrate, suppose that the common information collection responsibilities are assigned by contract to one of the parties. The purchasing party then runs a veracity risk: information may be filtered and possibly distorted to the advantage of the firm that has assumed the information collection responsibility. If checks are costly and proof of contractual violation difficult, contractual sharing arrangements manifestly experience short-run limitations. If, in
addition, small numbers prevail so that options are restricted, contractual sharing is subject to long-run risks as well. On this argument, observational economies are mainly to be attributed to strategic misrepresentation risks rather than to indivisibilities.

3. Convergence of Expectations

The issue to which the convergence of expectations argument is addressed is that, if there is a high degree of interdependence among successive stages of production and if occasions for adaptation are unpredictable yet common, coordinated responses may be difficult to secure if the separate stages are operated independently. March and Simon (1958) characterize the problem in the following terms:

Interdependence by itself does not cause difficulty if the pattern of interdependence is stable and fixed. For, in this case, each subprogram can be designed to take account of all the subprograms with which it interacts. Difficulties arise only if program execution rests on contingencies that cannot be predicted perfectly in advance. In this case, coordinating activity is required to secure agreement about the estimates that will be used as the basis for action, or to provide information to each subprogram unit about the activities of the others.

This reduces, in some respects, to a contractual incompleteness argument. Were it feasible exhaustively to stipulate the appropriate conditional responses, coordination could proceed by contract. This is ambitious, however; in the face of a highly variable and uncertain environment, the attempt to program responses is apt to be inefficient. To the extent that an unprogrammed (adaptive, sequential) decision process is employed instead, and in consideration of the severe incentive and control limitations that long-term contracts experience in these circumstances (See Part B above), vertical integration may be indicated.

But what of the possibility of short-term contracts? It is here that the convergence of expectations argument is of special importance. Thus assume that short-term contracts are not defective on account either of investment disincentives or first-mover advantages. It is Malmgren’s (1961) contention that such contracts may nevertheless be vitiated by the absence of structural constraints. The costs of negotiations and the time required to bring the system into adjustment by exclusive reliance on market (price) signals are apt to be great in relation to that which would obtain if successive states were integrated and administrative processes employed as well or instead.

E. Institutional Adaptations

Institutional adaptations of two types are distinguished: simple economic and extra-economic.

1. Simple Economic

As has been noted by others, vertical integration may be a device by which sales taxes on intermediate products are avoided, or a means by which to circumvent quota schemes and
price controls (Coase, 1937; Stigler, 1968). But vertical integration may also be undertaken because of the defective specification of property rights.

Although the appropriate assignment of property rights is a complex question, it reduces (equity considerations aside) to a simple criterion: What assignment yields maximum total product (Coase, 1960)? This depends jointly on imputation and negotiation expenses and on the incentives of the compensated party. So as to focus on the negotiation expense aspect, assume that imputation expenses are negligible and set the incentive question aside for the moment. An “appropriate” assignment of property rights will here be defined as one which automatically yields compensation in the amount of the external benefit or cost involved, while an “inappropriate” assignment is one that requires bargaining to bring the parties into adjustment. Thus if A and B are two parties and A’s activity imposes costs on B, the appropriate assignment of property rights is to require A to compensate B. If instead property rights were defined such that A is not required to compensate B, and assuming that the externality holds at the margin, efficient adaptation would occur only if B were to bribe A to bring his activity into adjustment—which entails bargaining. Only if the costs of such bargaining are neglected can the alternative specifications of property rights be said to be equivalent. For similar reasons, if A’s activity generates benefits for B, the appropriate specification of property rights will be to require B fully to compensate A. Harmonizing the otherwise divergent interests of the two parties by internalizing the transaction through vertical merger promises to overcome the haggling costs which result when property rights are left either undefined or inappropriately specified.

2. Other

Risk aversion refers to the degree of concavity in the utility valuation of pecuniary outcomes. Decision-makers who are risk averse will be concerned not merely with the expected value, but also with the dispersion in outcomes associated with alternative proposals: the greater the dispersion, the lower the utility valuation. . . . [D]ecision-makers who are the less risk averse will presumably assume the risk bearing function. Even, however, if attitudes toward risk were identical—in the sense that every individual (for any given set of initial endowments) would evaluate a proposal similarly—differing initial asset positions among the members of a population could warrant a specialization of the risk-bearing function, with possible firm and market structure effects (Knight, 1965).

Arrow calls attention to norms of social behavior, including ethical and moral codes. He observes in this connection that “It is useful for individuals to have some trust in each other’s word. In the absence of trust, it would become very costly to arrange for alternative sanctions and guarantees, and many opportunities for mutually beneficial cooperation would have to be foregone” (1969). One would expect, accordingly, that vertical integration would be more complete in a low-trust than a high-trust culture . . . .
III. CONCLUSIONS

That product markets have remarkable coordinating properties is, among economists at least, a secure proposition. That product markets are subject to failure in various respects and that internal organization may be substituted against the market in these circumstances is, if somewhat less familiar, scarcely novel. A systematic treatment of market failure as it bears on vertical integration, however, has not emerged.

Partly this is attributable to inattention to internal organization: the remarkable properties of firms that distinguish internal from market coordination have been neglected. But the fragmented nature of the market failure literature as it bears on vertical integration has also contributed to this condition; the extensive variety of circumstances in which internalization is attractive tends not to be fully appreciated.

The present effort attempts both to address the internal organization issue and to organize the market failure literature as it relates to vertical integration in a systematic way. The argument, however, by no means exhausts the issues that vertical integration raises. For one thing, the discussion of market failures may be incomplete in certain respects. For another, a parallel treatment of the sources and consequences of the failures of internal organization as they relate to vertical integration is needed. Third, the argument applies strictly to the vertical integration of production; although much of it may have equal relevance to backward vertical integration into raw materials and forward integration into distribution, it may have to be delimited in significant respects. Fourth, game theoretic considerations, which may permit the indicated indeterminacy of small numbers bargaining situations to be bounded, have been neglected. Finally, nothing in the present analysis establishes that observed degrees of vertical integration are not, from a social welfare standpoint, excessive. It should nevertheless be apparent that a broader a priori case for the vertical integration of production exists than is commonly acknowledged.


An outsider to the field of economics would probably take it for granted that economists have a highly developed theory of the firm. After all, firms are the engines of growth of modern capitalistic economies, and so economists must surely have fairly sophisticated views of how they behave. In fact, little could be further from the truth. Most formal models of the firm are extremely rudimentary, capable only of portraying hypothetical firms that bear little relation to the complex organizations we see in the world. Furthermore, theories that attempt to incorporate real world features of corporations, partnerships and the like often lack precision and rigor, and have therefore failed, by and large, to be accepted by the theoretical mainstream.

This Article attempts to give lawyers a sense of how economists think about firms. It does not pretend to offer a systematic survey of the area; rather, it highlights several ideas of particular importance, and then explores an alternative theoretical perspective from
which to view the firm. Part I introduces various established economic theories of the
firm. Part II turns to a newer theory of the firm, based not upon human capital structures,
but rather upon property rights. Part III synthesizes this property rights-based theory of
the firm with more established theories.

I. ESTABLISHED THEORIES

A. Neoclassical Theory

Any discussion of theories of the firm must start with the neoclassical approach, the
staple diet of modern economists. Developed over the last one hundred years or so, this
approach can be found in any modern-day textbook; in fact, in most textbooks, it is the
only theory of the firm presented.

Neoclassical theory views the firm as a set of feasible production plans. A manager
presides over this production set, buying and selling inputs and outputs in a spot market
and choosing the plan that maximizes owners’ welfare. Welfare is usually represented by
profit, or, if profit is uncertain so that profit-maximization is not well defined, by
expected net present value of future profit (possibly discounted for risk) or by market
value.

To many lawyers and economists, this is a caricature of the modern firm; it is rigorous
but rudimentary. At least three reasons help explain its prolonged survival. First, the
theory lends itself to an elegant and general mathematical formalization. Second, it is
very useful for analyzing how a firm’s production choices respond to exogenous change
in the environment, such as an increase in wages or a sales tax. Finally, the theory is also
very useful for analyzing the consequences of strategic interaction between firms under
conditions of imperfect competition; for example, it can help us understand the
relationship between the degree of concentration in an industry and that industry’s output
and price level.

Granted these strengths, neoclassical theory has some very clear weaknesses. It does not
explain how production is organized within a firm, how conflicts of interest between the
firm’s various constituencies—its owners, managers, workers, and consumers—are
resolved, or, more generally, how the goal of profit-maximization is achieved. More
subtly, neoclassical theory begs the question of what defines a given firm or what
determines its boundaries. Since the theory does not address the issue of each firm’s size
or extent, it does not explain the consequences of two firms choosing to merge, or of one
firm splitting itself into two or more smaller firms. Neoclassical theory describes in
rudimentary terms how firms function, but contributes little to any meaningful picture of
their structure.

B. Principal-Agent Theory

Principal-agent theory, an important development of the last fifteen years, addresses
some of the weaknesses of the neoclassical approach. Principal-agent theory recognizes
conflicts of interest between different economic actors, formalizing these conflicts
through the inclusion of observability problems and asymmetries of information. The
theory still views the firm as a production set, but now a professional manager makes production choices, such as investment or effort allocations, that the firm’s owners do not observe. Because the manager deals with the day-to-day operations of the firm, she also is presumed to have information about the firm’s profitability that the owners lack. In addition, the manager has other goals in mind beyond the owners’ welfare, such as on-the-job perks, an easy life, empire building, and so on. Under these conditions, principal-agent theory argues that it will be impossible for the owners to implement their own profit-maximizing plan directly, through a contract with the manager—in general, the owners will not even be able to tell ex post whether the manager has chosen the right plan. Instead, the owners will try to align the manager’s objectives with their own by putting the manager on an incentive scheme. Even under an optimal incentive scheme, however, the manager will put some weight on her own objectives at the expense of those of the owners, and conflicting interests remain. Hence, we have the beginnings of a managerial theory of the firm.

Principal-agent theory enriches neoclassical theory significantly, but still fails to answer the vital questions of what defines a firm and where the boundaries of its structure are located. To see why, consider the example of Fisher Body, which for many years has supplied car bodies to General Motors. Principal-agent theory can explain why it might make sense for GM and Fisher to write a profit-sharing agreement, whereby part of Fisher Body’s reward is based on GM’s profit from car sales: this encourages Fisher to supply high-quality inputs. The theory does not tell us, however, whether it matters if this profit-sharing agreement is accomplished through the merger of Fisher and GM into a single firm, with GM having authority over Fisher management; or whether GM and Fisher should remain as separate firms; or whether GM and Fisher should merge, with Fisher management having authority over GM management. In other words, principal-agent theory tells us about optimal incentive schemes, but not (at least directly) about organizational form. Hence, in the absence of a parallel between the two, which turns out to be difficult to draw, principal-agent theory provides no predictions about the nature and extent of the firm.

C. Transaction Cost Economics

While the neoclassical paradigm, modified by principal-agent theory, progressed along the above lines, a very different approach to the theory of the firm developed under the heading of transaction cost economics. Introduced in Coase’s famous 1937 article, transaction cost economics traces the existence of firms to the thinking, planning and contracting costs that accompany any transaction, costs usually ignored by the neoclassical paradigm. The idea is that in some situations these costs will be lower if a transaction is carried out within a firm rather than in the market. According to Coase, the main cost of transacting in the market is the cost of learning about and haggling over the terms of trade; this cost can be particularly large if the transaction is a long-term one in which learning and haggling must be performed repeatedly. Transaction costs can be reduced by giving one party authority over the terms of trade, at least within limits. But, according to Coase, this authority is precisely what defines a firm: within a firm, transactions occur as a result of instructions or orders issued by a boss, and the price mechanism is suppressed.
Such an arrangement, however, brings costs of its own. Concentrating authority in one person’s hands is likely to increase the cost of errors and lead to greater administrative rigidity. In Coase’s view, the boundaries of the firm occur at the point where the marginal cost savings from transacting within the firm equal these additional error and rigidity costs.

Coase’s ideas, although recognized as highly original, took a long time to catch on. There are probably two reasons for this. First, they remain to this day very hard to formalize. Second, there is a conceptual weakness, pointed out by Alchian and Demsetz, in the theory’s dichotomy between the role of authority within the firm and the role of consensual trade within the market. Consider, for example, Coase’s notion that an employer has authority over an employee—an employer can tell an employee what to do. Alchian and Demsetz questioned this, asking what ensures that the employee obeys the employer’s instructions. To put it another way, what happens to the employee if he disobeys these instructions? Will he be sued for breach of contract? Unlikely. Probably the worst that can happen is the employee will be fired. But firing is typically the sanction that one independent contractor will impose on another whose performance he does not like. To paraphrase Alchian and Demsetz’s criticism, it is not clear that an employer can tell an employee what to do, any more than a consumer can tell her grocer what to do (what vegetables to sell at what prices); in either case, a refusal will likely lead to a termination of the relationship, a firing. In the case of the grocer, this means that the consumer shops at another grocer. Thus, according to Alchian and Demsetz’s argument, Coase’s view that firms are characterized by authority relations does not really stand up.

Finding Coase’s characterization of the firm wanting, Alchian and Demsetz developed their own theory, based on joint production and monitoring. Transactions involving joint or team production require careful monitoring so that each actor’s contribution can be assessed. According to Alchian and Demsetz, the best way to provide the monitor with appropriate incentives is to give him the following bundle of rights, which effectively define ownership of the capitalist firm:

1) to be a residual claimant;
2) to observe input behavior;
3) to be the central party common to all contracts with inputs;
4) to alter membership of the team; and,
5) to sell rights 1-4.

We will return to some of these ideas below, but at this stage it suffices to note that the theory suffers from the same criticism levelled at Coase: it is unclear why the problems of joint production and monitoring must be solved through the firm and cannot be solved through the market. In fact, one does not need to look far to see examples of market solutions to these problems, such as auditing between independent contractors.
At the same time that doubts were being expressed about the specifics of Coase’s theory, Coase’s major idea—that firms arise to economize on transaction costs—was increasingly accepted. The exact nature of these transaction costs, however, remained unclear. What lay beyond the learning and haggling costs that, according to Coase, are a major component of market transactions? Professor Oliver Williamson has offered the deepest and most far-reaching analysis of these costs. Williamson recognized that transaction costs may assume particular importance in situations where economic actors make relationship-specific investments—investments to some extent specific to a particular set of individuals or assets. Examples of such investments include locating an electricity generating plant adjacent to a coal mine that is going to supply it; a firm’s expanding capacity to satisfy a particular customer’s demands; training a worker to operate a particular set of machines or to work with a particular group of individuals; or a worker’s relocating to a town where he has a new job.

In situations like these, there may be plenty of competition before the investments are made—there may be many coal mines next to which an electricity generating plant could locate or many towns to which a worker could move. But once the parties sink their investments, they are to some extent locked into each other. As a result, external markets will not provide a guide to the parties’ opportunity costs once the relationship is underway. This lack of information takes on great significance, since, in view of the size and degree of the specific investment, one would expect relationships like these to be long lasting.

In an ideal world, the lack of ex post market signals would pose no problem, since the parties could always write a long-term contract in advance of the investment, spelling out each agent’s obligations and the terms of the trade in every conceivable situation. In practice, however, thinking, negotiation and enforcement costs will usually make such a contract prohibitively expensive. As a result, parties must negotiate many of the terms of the relationship as they go along. Williamson argues that this leads to two sorts of costs. First, there will be costs associated with the ex post negotiation itself—the parties may engage in collectively wasteful activities to try to increase their own share of the ex post surplus; also, asymmetries of information may make some gains from trade difficult to realize. Second, and perhaps more fundamental, since a party’s bargaining power and resulting share of the ex post surplus may bear little relation to his ex ante investment, parties will have the wrong investment incentives at the ex ante stage. In particular, a far-sighted agent will choose her investment inefficiently from the point of view of her contracting partners, given that she realizes that these partners could expropriate part of her investment at the ex post stage.

In Williamson’s view, bringing a transaction from the market into the firm—the phenomenon of integration—mitigates this opportunistic behavior and improves investment incentives. Agent A is less likely to hold up agent B if A is an employee of B than if A is an independent contractor. However, Williamson does not spell out in precise terms the mechanism by which this reduction in opportunism occurs. Moreover, certain costs presumably accompany integration. Otherwise, all transactions would be carried out in firms, and the market would not be used at all. Williamson, however, leaves the precise nature of these costs unclear.
D. The Firm as a Nexus of Contracts

All the theories discussed so far suffer from the same weakness: while they throw light on the nature of contractual failure, none explains in a convincing or rigorous manner how bringing a transaction into the firm mitigates this failure.

One reaction to this weakness is to argue that it is not really a weakness at all. According to this point of view, the firm is simply a nexus of contracts, and there is therefore little point in trying to distinguish between transactions within a firm and those between firms. Rather, both categories of transactions are part of a continuum of types of contractual relations, with different firms or organizations representing different points on this continuum. In other words, each type of business organization represents nothing more than a particular “standard form” contract. One such “standard form” contract is a public corporation, characterized by limited liability, indefinite life, and freely transferable shares and votes. In principle it would be possible to create a contract with these characteristics each time it is needed, but, given that these characteristics are likely to be useful in many different contexts, it is much more convenient to be able to appeal to a “standard form.” Closely held corporations or partnerships are other examples of useful “standard forms.”

Viewing the firm as a nexus of contracts is helpful in drawing attention to the fact that contractual relations with employees, suppliers, customers, creditors and others are an essential aspect of the firm. Also, it serves to make it clear that the personalization of the firm implied by asking questions such as “what should be the objective function of the firm” . . . is seriously misleading. The firm is not an individual . . . . The “behavior” of the firm is like the behavior of a market, i.e., the outcome of a complex equilibrium process.

At the same time, the nexus of contracts approach does less to resolve the questions of what a firm is than to shift the terms of the debate. In particular, it leaves open the question of why particular “standard forms” are chosen. Perhaps more fundamentally, it begs the question of what limits the set of activities covered by a “standard form.” For example, corporations are characterized by limited liability, free transferability of shares, and indefinite life. But what limits the size of a corporation—what are the economic consequences of two corporations merging or of one corporation splitting itself into two? Given that mergers and breakups occur all the time, and at considerable transaction cost, it seems unlikely that such changes are cosmetic. Presumably they have some real effects on incentives and opportunistic behavior, but these effects remain unexplained.

II. A PROPERTY RIGHTS APPROACH TO THE FIRM

One way to resolve the question of how integration changes incentives is spelled out in recent literature that views the firm as a set of property rights. This approach is very much in the spirit of the transaction cost literature of Coase and Williamson, but differs
by focusing attention on the role of physical, that is, nonhuman, assets in a contractual relationship.

Consider an economic relationship of the type analyzed by Williamson, where relationship-specific investments are important and transaction costs make it impossible to write a comprehensive long-term contract to govern the terms of the relationship. Consider also the nonhuman assets that, in the postinvestment stage, make up this relationship. Given that the initial contract has gaps, missing provisions, or ambiguities, situations will typically occur in which some aspects of the use of these assets are not specified. For example, a contract between GM and Fisher might leave open certain aspects of maintenance policy for Fisher machines, or might not specify the speed of the production line or the number of shifts per day.

Take the position that the right to choose these missing aspects of usage resides with the owner of the asset. That is, ownership of an asset goes together with the possession of residual rights of control over that asset; the owner has the right to use the asset in any way not inconsistent with a prior contract, custom, or any law. Thus, the owner of Fisher assets would have the right to choose maintenance policy and production line speed to the extent that the initial contract was silent about these.

Finally, identify a firm with all the nonhuman assets that belong to it, assets that the firm’s owners possess by virtue of being owners of the firm. Included in this category are machines, inventories, buildings or locations, cash, client lists, patents, copyrights, and the rights and obligations embodied in outstanding contracts to the extent that these are also transferred with ownership. Human assets, however, are not included. Since human assets cannot be bought or sold, management and workers presumably own their human capital both before and after any merger.

We now have the basic ingredients of a theory of the firm. In a world of transaction costs and incomplete contracts, ex post residual rights of control will be important because, through their influence on asset usage, they will affect ex post bargaining power and the division of ex post surplus in a relationship. This division in turn will affect the incentives of actors to invest in that relationship. Hence, when contracts are incomplete, the boundaries of firms matter in that these boundaries determine who owns and controls which assets. In particular, a merger of two firms does not yield unambiguous benefits: to the extent that the (owner-)manager of the acquired firm loses control rights, his incentive to invest in the relationship will decrease. In addition, the shift in control may lower the investment incentives of workers in the acquired firm. In some cases these reductions in investment will be sufficiently great that nonintegration is preferable to integration.

Note that, according to this theory, when assessing the effects of integration, one must know not only the characteristics of the merging firms, but also who will own the merged company. If firms A and B integrate and A becomes the owner of the merged company, then A will presumably control the residual rights in the new firm. A can then use those rights to hold up the managers and workers of firm B. Should the situation be reversed, a different set of control relations would result in B exercising control over A, and A’s workers and managers would be liable to holdups by B.
It will be helpful to illustrate these ideas in the context of the Fisher Body-General Motors relationship. Suppose these companies have an initial contract that requires Fisher to supply GM with a certain number of car bodies each week. Imagine that demand for GM cars now rises and GM wants Fisher to increase the quantity it supplies. Suppose also that the initial contract is silent about this possibility, perhaps because of a difficulty in predicting Fisher’s costs of increasing supply. If Fisher is a separate company, GM presumably must secure Fisher’s permission to increase supply. That is, the status quo point in any contract renegotiation is where Fisher does not provide the extra bodies. In particular, GM does not have the right to go into Fisher’s factory and set the production line to supply the extra bodies; Fisher, as owner, has this residual right of control. The situation is very different if Fisher is a subdivision or subsidiary of GM, so that GM owns Fisher’s factory. In this case, if Fisher management refuses to supply the extra bodies, GM always has the option to fire management and hire someone else to supervise the factory and supply extra bodies (they could even run Fisher themselves on a temporary basis). The status quo point in the contract renegotiation is therefore quite different.

To put it very simply, if Fisher is a separate firm, Fisher management can threaten to make both Fisher assets and their own labor unavailable for the uncontracted-for supply increase. In contrast, if Fisher belongs to GM, Fisher management can only threaten to make their own labor unavailable. The latter threat will generally be much weaker than the former.

Although the status quo point in the contract renegotiation may depend on whether GM and Fisher are one firm rather than two, it does not follow that the outcomes after renegotiation will differ. In fact, if the benefits to GM of the extra car bodies exceed the costs to Fisher of supplying them, we might expect the parties to agree that the bodies should be supplied, regardless of the status quo point. However, the divisions of surplus in the two cases will be very different. If GM and Fisher are separate, GM may have to pay Fisher a large sum to persuade it to supply the extra bodies. In contrast, if GM owns Fisher’s plant, it may be able to enforce the extra supply at much lower cost since, as we have seen in this case, Fisher management has much reduced bargaining and threat power.

Anticipating the way surplus is divided, GM will typically be much more prepared to invest in machinery that is specifically geared to Fisher bodies if it owns Fisher than if Fisher is independent, since the threat of expropriation is reduced. The incentives for Fisher, however, may be quite the opposite. Fisher management will generally be much more willing to come up with cost-saving or quality-enhancing innovations if Fisher is an independent firm than if it is part of GM, because Fisher management is more likely to see a return on its activities. If Fisher is independent, it can extract some of GM’s surplus by threatening to deny GM access to the assets embodying these innovations. In contrast, if GM owns the assets, Fisher management faces total expropriation of the value of the innovation to the extent that the innovation is asset-specific rather than management-specific, and GM can threaten to hire a new management team to incorporate the innovation.
So far, we have discussed the effects of control changes on the incentives of top management. But workers’ incentives will also be affected. Consider, for example, the incentive of someone who works with Fisher assets to improve the quality of Fisher’s output by better learning some aspect of the production process. Suppose further that GM has a specific interest in this improvement in car body quality, and that none of Fisher’s other customers cares about it. There are many ways in which the worker might be rewarded for this, but one important reward is likely to come from the fact that the worker’s value to the Fisher-GM venture will rise in the future and, due to his additional skills, the worker will be able to extract some of these benefits through a higher wage or promotion. Note, however, that the worker’s ability to do this is greater if GM controls the assets than if Fisher does. In the former case, the worker will bargain directly with GM, the party that benefits from the worker’s increased skill. In the latter case, the worker will bargain with Fisher, who only receives a fraction of these benefits, since it must in turn bargain with GM to parlay these benefits into dollars. In consequence, the worker will typically capture a lower share of the surplus, and his incentive to make the improvement in the first place will fall.

In other words, given that the worker may be held up no matter who owns the Fisher assets—assuming that he, himself, does not—his incentives are greater if the number of possible hold-ups is smaller rather than larger. With Fisher management in control of the assets, there are two potential hold-ups: Fisher can deny the worker access to the assets, and GM can decline to pay more for the improved product. As a result, we might expect the worker to get, say, a third of his increased marginal product (supposing equal division with Fisher and GM). With GM management in control of the Fisher assets, there is only one potential hold-up, since the power to deny the worker his increased marginal product is concentrated in one agent’s hands. As a result, the worker in this case might be able to capture half of his increased marginal product (supposing equal division with GM).

The above reasoning applies to the case in which the improvement is specific to GM. Exactly the opposite conclusion would be reached, however, if the improvement were specific to Fisher, such as the worker learning how to reduce Fisher management’s costs of making car bodies, regardless of Fisher’s final customer (a cost reduction, furthermore, which could not be enjoyed by any substitute for Fisher management). In that event, the number of hold-ups is reduced by giving control of Fisher assets to Fisher management rather than GM. The reason is that with Fisher management in control, the worker bargains with the party who benefits directly from his increased productivity, whereas with GM management in control, he must bargain with an indirect recipient; GM must in turn bargain with Fisher management to benefit from the reduction in costs.

Up to this point we have assumed that GM management will control GM assets. This, however, need not be the case; in some situations it might make more sense for Fisher management to control these assets—for Fisher to buy up GM. One thing we can be sure of is that if GM and Fisher assets are sufficiently complementary, and initial contracts sufficiently incomplete, then the two sets of assets should be under common control. With extreme complementarity, no agent—whether manager or worker—can benefit from any increase in his marginal productivity unless he has access to both sets of assets (by the definition of extreme complementarity, each asset, by itself, is useless). Giving
control of these assets to two different management teams is therefore bound to be
detrimental to actors’ incentives, since it increases the number of parties with hold-up
power. This result confirms the notion that when lock-in effects are extreme, integration
will dominate nonintegration.

These ideas can be used to construct a theory of the firm’s boundaries. First, as we have
seen, highly complementary assets should be owned in common, which may provide a
minimum size for the firm. Second, as the firm grows beyond a certain point, the
manager at the center will become less and less important with regard to operations at the
periphery in the sense that increases in marginal product at the periphery are unlikely to
be specific either to this manager or to the assets at the center. At this stage, a new firm
should be created since giving the central manager control of the periphery will increase
hold-up problems without any compensating gains. It should also be clear from this line
of argument that, in the absence of significant lock-in effects, nonintegration is always
better than integration—it is optimal to do things through the market, for integration only
increases the number of potential hold-ups without any compensating gains.

Finally, it is worth noting that the property rights approach can explain how the purchase
of physical assets leads to control over human assets. To see this, consider again the GM-
Fisher hypothetical. We showed that someone working with Fisher assets is more likely
to improve Fisher’s output in a way that is specifically of value to GM if GM owns these
assets than if Fisher does. This result can be expressed more informally as follows: a
worker will put more weight on an actor’s objectives if that actor is the worker’s boss,
that is, if that actor controls the assets the worker works with, than otherwise. The
conclusion is quite Coasian in spirit, but the logic underlying it is very different. Coase
reaches this conclusion by assuming that a boss can tell a worker what to do; in contrast,
the property rights approach reaches it by showing that it is in a worker’s self-interest to
behave in this way, since it puts him in a stronger bargaining position with his boss later
on.

To put it slightly differently, the reason an employee is likely to be more responsive to
what his employer wants than a grocer is to what his customer wants is that the employer
has much more leverage over his employee than the customer has over his grocer. In
particular, the employer can deprive the employee of the assets he works with and hire
another employee to work with these assets, while the customer can only deprive the
grocer of his custom and as long as the customer is small, it is presumably not very
difficult for the grocer to find another customer.

III. PROPERTY RIGHTS AND THE ESTABLISHED THEORIES OF THE FIRM

The property rights approach has features in common with each of the approaches
described previously. It is based on maximizing behavior (like the neoclassical
approach); it emphasizes incentive issues (like the principal-agent approach); it
emphasizes contracting costs (like the transaction cost approach); it treats the firm as a
“standard form” contract (like the nexus of contracts approach); and, it relies on the idea
that a firm’s owner has the right to alter membership of the firm: the owner has the right
to decide who uses the firm’s assets and who doesn’t. Its advantage over these other
approaches, however, is its ability to explain both the costs and the benefits of integration; in particular, it shows how incentives change when one firm buys up another one.

Some react skeptically to the notion that a firm can be characterized completely by the nonhuman assets under its control. That is, there is a feeling that one should be able to make sense of a firm as a mode of organization, even if there are no definable assets on the scene. In his analysis of GM’s decision to acquire Fisher Body in 1926, Professor Klein argues that getting control over Fisher’s organizational assets rather than their physical capital was the crucial motivating factor:

By integrating with Fisher, General Motors acquired the Fisher Body organizational capital. This organization is embedded in the human capital of the employees at Fisher but is in some sense greater than the sum of its parts. The employees come and go but the organization maintains the memory of past trials and the knowledge of how to best do something (that is, how to make automobile bodies).

Klein’s conclusion is in no way inconsistent with the property rights approach. The control of physical capital can lead to control of human assets in the form of organizational capital. However, Klein appears to argue that his conclusion would hold true even if physical assets were irrelevant. The problem with this point of view is that, in the absence of physical assets, it is unclear how GM can get control over an intangible asset like organizational capital by purchasing Fisher. For example, what is to stop Fisher management from trying to reassert control of the organizational capital after the merger? Klein writes:

A threat that all the individuals will simultaneously shirk or leave if their wages were not increased to reflect the quasirents on the organizational capital generally will not be credible. After vertical integration the Fisher brothers will not be able to hold up General Motors by telling all the employees to leave General Motors and show up on Monday morning at a new address.

This conclusion is reasonable when physical capital is important since it would be difficult at best for Fisher employees to find a substitute for this capital, particularly by Monday morning. However, it is not reasonable in the absence of physical assets. In this case, to paraphrase Alchian and Demsetz, the Fisher brothers have no more ability to hold up GM by telling all the employees to leave GM or, more generally, by countermanding GM’s instructions, when Fisher is separate than when Fisher belongs to GM. Their ability to do so will be determined by factors such as the motivation, talent, knowledge and charisma of the Fisher brothers; the quality of worker information; and the degree of worker inertia—factors that do not seem to have anything to do with ownership structure. To put it another way, GM’s response to a hold-up attempt by the Fisher brothers will be the same whether GM owns Fisher or Fisher is independent: to try to persuade Fisher workers to desert the Fisher brothers and join GM.
As noted previously, one of the weaknesses of the property rights approach as described here is that it does not take account of the separation of ownership and control present in large, publicly held corporations. In principle, it should be possible to extend the existing analysis to such situations. A public corporation can still be usefully considered a collection of assets, with ownership providing control rights over these assets. Now, however, the picture is more complicated. Although owners (shareholders) typically retain some control rights, such as the right to replace the board of directors, in practice they delegate many others to management, at least on a day-to-day basis. In addition, some of the shareholders’ rights shift to creditors during periods of financial distress. Developing a formal model of the firm that contains all these features, and that includes also an explanation of the firm’s financial structure, is an important and challenging task for future research. Fortunately, recent work suggests that the task is not an impossible one. . . .
§ 6. Economics and the Public-Private Choice

§ 6.1. Public Ownership as Political Management


The second fundamental idea that the privatizers subscribed to was that, at least in Russia, political influence over economic life was the fundamental cause of economic inefficiency, and that the principal objective of reform was, therefore, to depoliticize economic life. Price liberalization fosters depoliticization because it deprives politicians of the opportunity to allocate goods. Privatization fosters depoliticization because it robs politicians of control over firms. When firms are subject to political influence, they cater to politicians by producing goods that consumers do not want, employing too many people, delivering output below cost to buyers favored by politicians and engaging in other grossly inefficient practices. Under socialism, firms were forced to pursue these political objectives, whereas during the Gorbachev transition, they pursued them in exchange for subsidies. The goal of privatization was to sever the links between enterprise managers and politicians, including both the Moscow industrial ministers and local officials, so as to force firms to cater to consumers and shareholders rather than politicians. There was no other way to achieve restructuring and efficient operation of firms. . . . In retrospect, it is remarkable how clearly the privatizers agreed on the overwhelming need for depoliticization. To a large extent, Russian privatization succeeded because the privatizers focused on this primary objective and were prepared to compromise on less important issues.

Again, the idea that depoliticization is the principal goal of privatization was not entirely new. For example, British Chancellor of the Exchequer Nigel Lawson in his memoirs stated clearly that this was the principal objective of British privatization in the 1980s. At the same time, many people believed, and still believe, that other objectives are as, or more, important. For example, many investment bankers, representatives of international lending agencies, and business people believed that Russian privatization should focus on selling firms one by one to substantial investors, especially foreign ones. Advocates of this view also often favored restructuring firms and even whole industries prior to privatization so as to prepare them for private ownership. The case-by-case method of privatization had an added advantage of generating revenue for the government. The Russian government, of course, badly needed cash. Thus to many Russians, as well as Westerners, case-by-case privatization appeared to be the right approach.

Nonetheless, if political influence is the main cause of inefficiency, then it is absurd to trust the ministries, or the investment bankers they designate, with finding foreign partners or with restructuring firms and entire industries. Russia’s privatizers believed
that politicians would use their power not to enhance efficiency, but to expand their economically undesirable control. As a result, the privatizers never seriously considered case-by-case privatization, and excluded those who advocated it too aggressively from the process.

An alternative view, commonly articulated by academics working on privatization, stated that the goal of privatization should be to create a corporate governance system that would alleviate the conflicts of interest between managers and shareholders. In this view, privatization needs to solve the problem of managerial discretion rather than that of political discretion. The Russian privatizers tried hard to deal with this problem, but the goal of a better corporate governance system was clearly subordinate to that of depoliticization. In fact, the Russian privatizers paid a great deal of attention to Poland, where the focus on corporate governance gave rise to a privatization scheme in which firms were to be controlled by large, government-sponsored mutual funds. This scheme was rejected for Russia precisely because of the expectation that any mutual fund created by the government would be politically influenced. It is worth noting that Polish privatization never got off the ground. Perhaps Russia avoided this failure by focusing on the right problem.


Everywhere else in this book, we assume that firm behavior is motivated by the desire to maximize profit. Profit maximization seems to be the single most plausible objective of a privately owned firm, as it results in the maximization of shareholders’ wealth. By specifying the objective of a firm (or, more generally, an economic agent), we can make predictions as to how it will respond to government policies. Such predictions are essential in determining the effects of those policies. For example, the Averch-Johnson effect tells us that rate-of-return regulation results in a profit-maximizing firm overcapitalizing. By assuming that a firm maximizes profit subject to rate-of-return regulation, we are then able to identify an inefficiency resulting from government policy.

To compare a policy of regulation with one of public enterprise, we need to provide a theory as to how a public enterprise behaves. Because it is not privately owned, it is unlikely to maximize profit. It would be nice if a public enterprise was to maximize social welfare, but it would be hasty to jump to the conclusion that it does so. This is the difference between positive and normative economics. From a normative perspective, a public enterprise designed to handle the natural monopoly problem should pursue social welfare maximization. A positive theory tells us, in contrast, how a public enterprise actually does indeed behave. In order to derive such a theory, we will present a more general model of the firm than has been previously discussed.
MANAGERIAL MODEL OF A FIRM

An element common to almost all large organizations is a separation of ownership and managerial control. The modern private corporation is a classic example. Ownership is held by many shareholders, each with typically a small or negligible percentage of shares. As a result, day-to-day control of the firm rests with the appointed manager, though ultimate control lies with the shareholders. This statement is analogously true for most public enterprises, as ownership rests in the populace of the relevant political jurisdiction but control lies in the manager appointed by an elected official or a government bureaucrat, whichever it may be. In today’s economy it is fair to say that the separation of ownership and control is indicative of almost all large organizations.

Let us construct a model of the firm based upon the recognition that managerial control does not rest with the owners but rather with a paid employee. If owners have as much information as managers and can perfectly monitor the actions of the manager, this separation issue is of no importance. In that case, if the manager does not do exactly what is in the owners’ best interest, she will be fired. In spite of the separation of ownership and control, the results are the same as when the organization is owner-operated.

Of course, it is quite unreasonable to assume that owners can even come close to perfectly monitoring the manager. It is generally not in the best interests of any individual owner to spend the resources required to closely observe the manager’s actions. Furthermore, through her day-to-day intimacy with the operations of the firm, the manager has much better information than the owners. Thus, even if the owners did observe the actions of the manager (for example, in choosing project A over project B), it is difficult for the owners to determine whether the manager is acting in their best interests.

These two conditions—the separation of ownership and control and the imperfect monitoring of managerial behavior—are characteristic of most major private and public enterprises. The managerial model of the firm has been developed to describe how such an organization behaves. A central assumption of this model is that the manager acts to maximize her own utility subject to the constraints and incentives instituted by the owners. Within this model, the goal of the owners is to construct a set of constraints and incentives so as to induce the manager to act in their best interests. Of course, the goal of the manager is to maximize her own utility. Our goal is to derive predictions as to how such a firm behaves.

In terms of her relationship with the firm, a manager’s utility depends on income, nonpecuniary benefits, and effort. The manager’s utility is greater, the greater is income and the greater are nonpecuniary benefits. The latter may include such items as the prestige of having a large staff or having a company jet at her disposal. In contrast, utility is less, the more effort the manager exerts. We can think of more effort as entailing less leisure, and less leisure generally reduces utility.

Not being perfectly monitored, the manager is likely to choose less effort and more nonpecuniary benefits than are in the owners’ best interests. This is why the owners must
design incentive schemes and create constraints in order to induce the manager to work hard and act efficiently (note that choosing to have a large staff for the purpose of prestige can be inefficient because it may not sufficiently increase productivity so as to warrant its cost).

Thus far we have not had to specify whether the firm is a private or a public enterprise because the manager in both organizations will tend to act the same . . . . The differences in the two organizations lie in (1) the interests of the owners and (2) the mechanisms available to induce the manager to act in the owners’ interests.

**Managerial Model of a Private Enterprise**

In a private enterprise, owners want the manager to act so as to maximize profit. A key mechanism to induce such behavior is to make the manager’s financial compensation depend on the profitability of the firm, that is, the use of an incentive scheme in which income increases with firm profit. This often takes the form of bonuses or issuing stock and stock options to the manager so as to make her a shareholder. Of course, this hardly solves the problem, because the manager knows that if she works harder and increases profit, she will get only a small part of that profit increase. Making income depend on firm profit provides the right incentives, but it cannot entirely solve the problem of the separation of ownership and control.

An implicit incentive scheme also arises in the labor market. If a manager’s hard work results in the firm earning a high rate of return, this will increase her reputation as a productive manager. Her value in the labor market will rise, resulting in higher future income. Thus, not only does superior performance increase current income, through bonuses and stock holdings, but future income is increased as well, by enhancing the manager’s reputation.

A final mechanism that induces the manager of a private enterprise to maximize profit is the threat of being fired if the firm performs poorly under her management. If owners observe poor performance, they can pursue actions to have the manager dismissed. Such a dismissal reduces not only current income but also future earnings, by causing deterioration of her reputation in the labor market. In reality, this is not a very effective constraint on managerial behavior because it is unlikely to be used. Given the wide dispersal of ownership in most privately owned firms, the owners are not usually in a position to credibly threaten such a response.

Nevertheless, the threat of being fired can arise quite effectively via the capital market. A firm that is being run inefficiently presents a profitable opportunity for investors. Because it is operating below its potential, the shares of the firm are selling at a relatively cheap price. Thus, investors (or, in today’s parlance, raiders) can purchase shares at this low price, install new and efficient management, and return the firm to efficiency. The increase in profit earned by the firm is the financial return to the investors. In some cases, the investors immediately resell the firm at the higher share price. Thus, even if the ownership of a firm is widely dispersed, poor firm performance will induce investors to buy shares and concentrate ownership in order to replace ineffective management.
Managerial Model of a Public Enterprise

Is there a similarly strong case to be made for the manager of a public enterprise to act in the owners’ best interests and thus maximize social welfare? For two reasons, this is unlikely to be true. First, welfare, unlike profit, is difficult to measure. The use of imperfect indicators should provide opportunities for the manager to act in his own interests and not those of society. Second, the constraints imposed on the manager of a private enterprise by the capital market are conspicuously absent. This lack of constraints will give the manager of a public enterprise greater discretion in his actions.

Although welfare is difficult to measure, there are substitutes that elected officials and voters can use to gauge the performance of a public enterprise. One obviously is price. A second set of attributes of output, for example, reliability of the service provided by the public enterprise. For the case of an electric utility, frequent blackouts would signal poor performance.

Several theories of public enterprise behavior have been formulated on the premise that managers act to maximize political support and this goal is achieved by producing those attributes that signal good performance. By increasing political support, a manager raises his income and the likelihood of increased job tenure. In “A Theory of Government Enterprise,” Cotton Lindsay notes,

[M]anagers are therefore influenced to divert resources from the production of attributes which will not be monitored to those which will. In so doing they will increase the perceived value of the [public enterprise’s] output.

For example, suppose that the quality of a service is easily observable but cost is not. The manager will then tend to choose too high a level of quality in order to increase political support. The important point is that the imperfections inherent in using attributes to measure managerial performance are likely to lead to strategic behavior by the manager. The manager will tend to produce those attributes to a degree that is excessive from society’s perspective.

Sam Peltzman proposed a different theory, in which the manager of a public enterprise uses price to maximize political support. This theory yields several interesting results. Let us suppose that political/voter support is greater, the smaller is the subsidy given to the public enterprise. This assumption is reasonable because the higher the subsidy, the more tax revenue must be raised, and voters dislike tax increases. If we assume that the public enterprise must earn a normal rate of return if it is to raise new capital, then the subsidy will equal $-\pi(P)$, where $\pi(P)$ is profit for a given price. If $\pi(P) > 0$, then the subsidy is negative, as the public enterprise adds to government revenues. Let us also assume that voter support is greater at lower prices, inasmuch as voters are also consumers.

One prediction of this model is that a public enterprise’s manager will set a price below that which maximizes profit. To show why this is true, consider price being set at the profit-maximizing level, which we denote as $P^m$. Recall that marginal profit (that is, the change in profit from a small change in price) equals zero at a price of $P^m$. The reason is
that if marginal profit was positive, then profit could be increased by raising price, which
contradicts $P^m$ being the profit-maximizing price. Similarly, if marginal profit was
negative at $P^m$, then profit could be increased by lowering price. Hence, marginal profit
must be zero at $P^m$.

Now consider the public enterprise charging a price slightly less than $P^m$. There is a direct
and an indirect effect on political support. The direct effect is that voters care about price
in their role as consumers. The indirect effect is that price affects firm profit, which
affects the amount of the government subsidy. Voters care about the amount of the
subsidy, as they are also taxpayers. Because marginal profit at $P^m$ is zero, a very slight
reduction in price does not affect firm profit, so that it does not affect the amount of the
subsidy. The indirect effect on political support from charging a price slightly less than
$P^m$ is then zero. On the other hand, the direct effect of charging a lower price raises
political support, as voters always like lower prices . . . . It follows that a public enterprise
manager’s political support is increased by setting a price below that which maximizes
profit. We conclude that a public enterprise will charge a lower price than a private
enterprise. . . .

A second result is that price will be lower for voters than nonvoters. Consider a public
enterprise located in political jurisdiction A but serving residents of both A and
jurisdiction B. The owners of the public firm are just the voters in A. By setting a higher
price for consumers in B, the manager can increase profit and reduce taxes, thereby
raising political support. This purpose is achieved without having to increase the price
charged to voters (that is, consumers in A). Of course, there is an upper bound to the
price that can be charged to residents of B, as they can always decide to form their own
public enterprise or having a private enterprise supply them.

A third prediction from the Peltzman model is that a public enterprise will pursue less
price discrimination than a private enterprise in order to provide a more uniform
treatment of consumers. Differential treatment can reduce political support because it
alienates certain voters. Thus, relative to a private enterprise, one would expect less
discrimination between industrial and residential users of the service as well as less
discrimination over time-of-day usage.

Although managers certainly desire political support, it is an overly narrow perspective to
think that they act simply so as to maximize political support. Managers also value
nonpecuniary benefits and leisure, not just income. For the case of a private enterprise,
we argued that the capital market is an important force in constraining the manager in her
desire to divert resources to nonpecuniary benefits and to shirk in performing her duties.
The threat of the firm’s being acquired and current management’s being replaced can be
an effective deterrent against the manager of a private enterprise’s acting too
inefficiently.

Perhaps the most important difference between public and private enterprise is that there
is no comparable disciplining force faced by the manager of a public enterprise. The
reason is that there is no feasible mechanism by which the ownership of a public
enterprise can be transferred. An owner of a public firm can forgo his share of ownership
only by moving to another political jurisdiction. Hence, there is no comparable mechanism to concentrate ownership as occurs in the capital market. The fact that owners can shift their ownership to another jurisdiction in response to poor performance is unlikely anyway. The performance of public enterprise is just one factor among many that influence an individual’s decision as to where to live.

The significance of the nontransferability of ownership is that it gives the manager of a public enterprise considerably more discretion than his private enterprise counterpart in pursuing his own personal interests. To make his work environment more pleasant, a manager may choose to minimize labor strife by providing higher wages. Similarly, he may choose to overinvest in capacity to prevent potential shortages and consumer strife. It has also been suggested that a manager will choose simpler pricing schedules and adjust them less frequently than a manager of a private enterprise. All these actions tend to increase nonpecuniary benefits from the job.

**COMPARISON OF PUBLIC AND PRIVATE ENTERPRISE**

Relative to an unregulated private enterprise, our analysis suggests that a public enterprise will price lower, practice less price discrimination, and earn lower profits. There also seems to be a plausible case for public enterprise to be less efficient. A manager of a public enterprise has a tendency to use more capital and labor in order to reap nonpecuniary benefits, such as fewer consumer complaints and an absence of labor strife.

These differences in behavior are generated by two factors. First, income and job tenure are raised by increasing firm profit for the manager of a private enterprise but are raised by increasing political support for a public enterprise’s manager. This difference results in lower price for a public enterprise and greater inefficiency, for example, overinvesting in product reliability to gain consumer-voter support. The second factor is that the manager of a private enterprise is threatened by the disciplining force of the capital market. If she performs poorly, there is a mechanism by which to replace her even though ownership may be widely dispersed. Because of the nontransferability of ownership in the case of a public enterprise, there is no similar constraint on the manager of a public enterprise. She then has greater discretion to use resources to maximize her own utility rather than social welfare.

Of greater relevance to our objective, however, is the comparison between a public enterprise and a regulated private enterprise. [In the surrounding chapters, the authors have been discussing “natural monopolies,” such as electric utilities, that are subject to rate-of-return regulation.] How this comparison differs from the preceding depends very much on the effectiveness of the regulatory constraint. If regulation is not very binding on the firm, then the preceding analysis applies. However, suppose that the regulatory constraint is binding. We know that regulation will cause price and profit to be lower, just as we found for a public enterprise. Whether price and profit differ much between the two institutions is then an empirical question. . . .
In contrast, there is no reason to expect a regulated private enterprise not to practice price discrimination extensively. This is a potential source of difference from a public enterprise. We know by the Averch-Johnson effect that a private enterprise that is subject to rate-of-return regulation also has a tendency to overcapitalize. However, it still does not have the discretion to act inefficiently, like a public enterprise. Even for a regulated private enterprise, the capital market presents a constraining force on inefficient behavior. Furthermore, any increase in cost between rate hearings will reduce firm profit. Thus, one would expect a regulated private enterprise to be more efficient than a public enterprise.

According to this analysis, both a regulated private enterprise and a public enterprise will experience productive inefficiencies. Rate-of-return regulation results in the private firm overcapitalizing, whereas the public firm has less incentive to act efficiently inasmuch as its manager is interested more in political support than profit. From a productive efficiency standpoint, which approach to the natural monopoly problem is preferred is then an empirical question. In the next section we will examine how regulation and public enterprise have performed in the provision of electricity in municipalities.

[The authors then compare municipal and private electric utilities by analyzing studies of pricing behavior, allocative efficiency, and productive efficiency. They conclude that the evidence on relative efficiency is “mixed. Nevertheless, a survey of comparative studies of industries ranging from electric utilities to refuse collection to weather forecasting provides general support for the hypothesis that there is greater productive efficiency with private enterprise. Most of these studies conclude that publicly owned firms are less efficient than privately owned firms.” The authors also present a comparative study of a public and a private airline in Australia concluding that the private airline was more efficient.]

§ 6.2. Public Ownership and the Theory of the Firm

§ 6.2.1. Andrei Shleifer, State Versus Private Ownership, J. ECON. PERSPECTIVES, Fall 1998, at 133

What kinds of goods and services should be provided by government employees as opposed to private firms? Should government workers make steel and cars in government-owned factories? Should teachers and doctors be publicly employed or should they work for private schools and practices? Should garbage be picked up by civil servants or employees of private garbage haulers? Should the whole economy be “socialized”? Although these are age-old questions in economics, the answers economists give to them, as well as the reasons for arriving at these answers, have been changing. In this paper, I describe some recent ways of thinking about government ownership.

---

5...[W]hen a regulated firm overcapitalizes, it does so in order to maximize profits by increasing its rate base. Although the capital-labor ratio is inefficient from a production perspective, the resulting ratio is the one that maximizes profits.
Half a century ago, economists were quick to favor government ownership of firms as soon as any market inequities or imperfections, such as monopoly power or externalities, were even suspected. Thus Arthur Lewis (1949), concerned with monopoly power, advocated the nationalization of land, mineral deposits, telephone service, insurance, and the motor car industry. For similar reasons, James Meade (1948) favored “socialization” of the iron and steel, as well as the chemical, industries. Maurice Allais (1947), always a step ahead of his English-speaking peers, argued for nationalization of a few firms in each (!) industry to facilitate the comparison of public and private ownership. At that time, privatization of such services as incarceration and education was evidently not even discussed by serious scholars.

These comments by future Nobel laureates were part of a broader debate over capitalism, socialism and the role of planning in a market economy, which raged in the 1930s and 1940s. The views of serious economists ranged from advocacy of socialism (Lange, 1936; Lerner, 1944), to fierce opposition to socialism (Hayek, 1944; Jewkes, 1948), to a reluctant concession that socialism is bad but inevitable because capitalism was running out of steam (Schumpeter, 1942). A remarkable aspect of this debate is that even many of the laissez-faire economists focused overwhelmingly on the goal of achieving competitive prices, even at the cost of accepting government ownership in non-competitive industries. Thus Henry Simons (1934), in “A Positive Program for Laissez Faire,” writes: “The state should face the necessity of actually taking over, owning, and managing directly, both the railroads and the utilities, and all other industries in which it is impossible to maintain effectively competitive conditions.” Simons’s advice is partly a response to the failures of regulation during the Depression, but the acquiescence of this libertarian economist to public ownership is symptomatic of the professional sentiment of the times. Pigou (1938), Schumpeter (1942) and Robbins (1947) were not too opposed to state ownership either, although Pigou recognized most clearly the dangers of bureaucratic control. In the first edition of his *Economics* text, Samuelson (1948) was more skeptical of socialism, writing: “It is only too easy to gloss over the tremendous dynamic vitality of our mixed free enterprise system, which, with all its faults, has given the world a century of progress such as an actual socialized order might find it impossible to equal.” Even Samuelson, however, focuses exclusively on the allocative role of prices, and does not say anything about ownership.

Consistent with the evident lack of aversion to state ownership, the postwar state assumed an enormous role in production throughout the world, owning everything from land and mines to industrial factories and communications to banks and insurance companies to hospitals and schools—even in market economies! In some of these economies, such as Japan, the United States, and Germany, government ownership was restrained, while in others, such as Italy, France, and Austria, the state assumed control of significant parts of production. Most developing countries opted for state ownership of the so-called “strategic” sectors. In socialist economies, the state came to own not just the strategic sectors, but everything else as well.

How the world has changed! In the last 20 years, governments in market economies throughout the world have privatized the very state firms in steel, energy,
telecommunications and financial services that the Nobel laureates approvingly saw nationalized a few decades earlier. Communism has collapsed almost everywhere in the world, and reform governments throughout the formerly socialist world have embarked on massive privatization programs. The economic policies of developing countries turned squarely to private ownership. In market economies, government provision of such basic services as garbage collection and education has come into question, and has increasingly been replaced by private provision, though still paid for largely from tax revenues.

Although some early voices, most notably that of Milton Friedman (1962), did rise to oppose state ownership, it is fair to say that postwar economists generally failed to anticipate its grotesque failure. In recent years, however, the evidence on the failures of state ownership in economies around the world has begun to accumulate (World Bank, 1995), and advances in the theories of ownership and contracting have reopened the question of state versus private provision. The contracting perspective distinguishes sharply between the government paying for a particular service, such as education, and providing it in-house. This perspective also helps to identify the opportunities for achieving social goals through private supply by a firm that may operate under a government contract or regulation. In a sense, the issues here are closely related to the vertical integration literature (Coase, 1937), except the question is that of the “make or buy” decision by the government rather than by a private firm.

When the opportunities for government contracting are exploited, the benefits of outright state ownership become elusive, even when social goals are taken into account. Moreover, it becomes clear that private ownership is the crucial source of incentives to innovate and become efficient, which accounts for what Samuelson (1948) called the “tremendous vitality” of the free enterprise system. The contractual perspective can serve as the basis of a theory of optimal provision when the government maximizes social welfare. It also allows us to think about an imperfect government, which maximizes political goals such as patronage or simply the income of politicians through bribes. Generally speaking, the case against state ownership becomes stronger when political considerations enter the calculation.

This paper begins by evaluating the case for in-house provision of goods and services by employees of a benevolent government. It argues that the conditions under which government ownership is superior in a country with good contract enforcement are very limited, and involve particular cases where soft incentives are extremely valuable and competition is very limited. I then turn to the more realistic case of a non-benevolent government, which helps to explain why the gains from privatization in many instances have been so enormous. Elementary and secondary education offer a particularly vivid example where I believe the case for near-monopoly government provision in an advanced democracy is indefensible.

The bottom line of this paper is simple. The 1940s case that government production is broadly desirable is no longer convincing. This case was motivated in part by such

---

1 In the prewar and predepression years, economists were a good deal more skeptical about state ownership. A truly brilliant statement of this skepticism, to which I return below, is by Alfred Marshall (1907).
empirical observations as some successes of government control during the war, the failures of competition and regulation during the Great Depression, and the apparent successes of Soviet industrialization, but also by a misunderstanding of the consequences of political control of firms, and by a substantial disregard of the importance of innovation in market economies. Today, the war and the Depression are no longer as vivid, and the communist economies have collapsed. As importantly, the quality of contracting and regulation have improved, competition has become more effective, the dangers of politicization of production have become self-evident, and the appreciation of the innovative potential of entrepreneurial firms is at a new high. For all these reasons, the benefits of reducing the role of government as a producer are becoming apparent and are beginning to be exploited.

THE DESIRABLE SCOPE OF A BENEVOLENT GOVERNMENT

Suppose that the government wants to have a good or service delivered to some consumers. The product can be food or shelter, steel or phone service, education, health care or incarceration. The government might wish to pay for some of this good and service out of its budget, or it may have views on the characteristics of this good, such as the price, even though the consumers buy it on their own. Should the government hire its own employees to deliver the service, or should it relinquish the provision to a (possibly regulated) private supplier? Does the mode of provision matter even when the government pays? In this section, I examine these questions for the case of a benevolent government.

The first key insight of the contracting approach deals with certain ways in which it does not matter whether the activity is provided in-house or contracted out. If the government knows exactly what it wants the producer to make, then it can put its wishes into the contract (or a regulation) and enforce this contract. In this case, the difference between in-house provision and contracting out disappears. Take some obvious cases where this point is crucial. A common argument for government ownership of the postal service is to enable the government to force the delivery of mail to sparsely populated areas, where it would be unprofitable to deliver it privately (Tierney, 1988). From the contractual perspective, this argument is weak. The government can always bind private companies that compete for a mail delivery concession to go wherever the government wants, or it can alternatively regulate these companies when entry is free. It cannot be so difficult to write the appropriate contract or regulation; after all, the government now tells the U.S. Postal Service where it wants the mail to be delivered.

Or take monopoly power, and the need for pricing restraint, as the argument for state ownership. If the government can describe the products that the monopoly delivers, it can always regulate a private monopoly, and often has done so in the cases of utilities or telecommunications. Once regulation is allowed, the 1940s case for superiority of direct government ownership evaporates. Or take, finally, the common case against private provision of schooling that holds that private schools will avoid taking problem students, whom public schools can be compelled to take. Again, if the government pays for at least some of the education through allocating vouchers to consumers, and can make universal access be a condition for redeeming a school’s vouchers, the problem can to a first
approximation be addressed contractually. With perfect contracting and regulation, there is no difference between state and private provision of goods and services. Of course, this result is not an endorsement of any and all regulation, but only a statement that the pursuit of “social goals” does not, on its own, make the case for government ownership.

The difference between provision modes reappears once we recognize that, in many instances, contracting opportunities are limited, and the government cannot fully anticipate, describe, stipulate, regulate and enforce exactly what it wants. A particular way to describe such limited contracting opportunities is the theory of incomplete contracts as developed by Grossman and Hart (1986), Hart and Moore (1990), and Hart (1995). These authors focus generally (and not just to model government) on contractual incompleteness, and more specifically on the idea that ownership of assets gives the owner control and bargaining power in situations where contracts do not specify what has to be done. As a consequence, ownership strengthens the owner’s incentives to make investments that improve the ways or reduce the costs of using the assets, because the owner has the power to reap more of the rewards on these innovations. An owner of a postal business who invents a better way to deliver mail can implement this innovation and profit from it. In contrast, if the government or someone else owns the business, the inventor needs the agreement of the owner to implement the innovation, and thus must share the benefits of the invention with this owner. Without the bargaining chip provided by ownership, the incentives to invest and innovate are lower.

Hart, Shleifer and Vishny (1997) apply Grossman-Hart-Moore theory to the choice between private and public provision. They broadly refer to the characteristics of the product with respect to which a contract is incomplete as non-contractible “quality.” This quality can stand for how well prisons treat inmates, how clean utilities keep the water, how well schools inculcate patriotism in their students, how long it takes a letter to get to a remote area, or how innovative car makers are. The choice of public versus private provision depends on how different ownership patterns affect the incentives to deliver this non-contractible quality, as well as on the cost of such delivery. The efficiency concept here is intended to incorporate fully the social value of quality.

To focus on both cost efficiency and quality, Hart, Shleifer and Vishny (1997) consider two types of investment incentives: those to reduce costs and those to improve quality or innovate. When assets are publicly owned, the public manager has relatively weak incentives to make either of these investments, because this manager is not the owner and hence gets only a fraction of the return. In contrast, private regulated contractors have much stronger incentives because, as owners, they get more of the returns on the investment. Which ownership structure is more efficient depends on whether having high-powered incentives to invest and innovate is a good idea.²

² The emphasis on soft incentives of government employees is common to many studies of private versus public bureaucracies, including Stiglitz (1989, 1994), Wilson (1989), Holmström and Milgrom (1991, 1994), Tirole (1994), and Williamson (1994). In all these studies, however, the conclusion of soft incentives of government employees is derived from assumptions made in addition to that of government ownership of assets, such as civil service rules, legal restrictions on compensation of bureaucrats,
The weak incentives of government employees with respect to both cost reduction and quality innovation underlie the basic case for the superiority of private ownership, a case that has been confirmed by the variety of empirical studies and general observation. It does not make sense for the government to own firms—be they steel mills, airlines, or grocery stores—when private firms can deliver the same, or even superior, quality of goods at a lower cost. Indeed, with few exceptions, the evidence on privatization points to both cost reductions and quality improvements in private relative to public firms (World Bank, 1995). For example, Barberis et al. (1996) find that both quality and efficiency improve in privatized shops in Russia, and Ehrlich et al. (1994) find higher productivity growth in private than in state airlines. Contrary to Schumpeter’s (1942) fears, private entrepreneurship remains a locomotive of progress.

The case for focusing on innovation in assessing the relative merits of government and private provision was made brilliantly by Marshall (1907). Having explained that innovation accounts for the enormous increase in British living standards in the 19th century, Marshall turned to the question of socialism, and argued that the government is generally a poor innovator: “A Government could print a good edition of Shakespeare’s works, but it could not get them written . . . . Every new extension of Governmental work in branches of production which need ceaseless creation and initiative is to be regarded as prima facie anti-social, because it retards the growth of that knowledge and those ideas which are incomparably the most important form of collective wealth.” The underappreciation of the importance of innovation has bedeviled most advocates of government ownership, including the Nobel laureates I quoted.

There is, however, a class of cases where the argument against government ownership is not as straightforward. In these cases, cost reductions for which private suppliers have stronger incentives have potentially deleterious effects on the non-contractible quality. For example, private prisons might abuse prisoners by hiring cheaper guards and failing to train them, private hospitals may refuse to treat patients on whom hospitals generally lose money, private schools might substitute less effective teachers’ aides for more expensive teachers, and so on. In such situations, strong incentives may lead to inefficient outcomes or, put differently, the efficient producer might need to have soft incentives. Ironically, the government sometimes becomes the efficient producer precisely because its employees are not motivated to find ways of holding costs down.

The modern case for government ownership can often be seen from precisely this perspective. Advocates of such ownership want to have state prisons so as to avoid untrained low-wage guards, state water utilities to force investment in purification, and state car makers to make them invest in environmentally friendly products. As it turns out, however, this case for state ownership must be made carefully, and even in most of the situations where cost reduction has adverse consequences for non-contractible quality, private ownership is still superior.

An advantage of the Grossman-Hart-Moore approach is that the soft incentives are derived from the very fact of government ownership rather than from these additional assumptions.
To begin, even when cost reduction does have an adverse effect on non-contractible quality, the greater incentives of private contractors to innovate may still shift the efficiency balance in their favor. For example, if the government is interested in a design and procurement of a new weapons system, it may prefer to hire a private contractor despite the concern that this contractor would sacrifice some aspects of quality to reduce costs (for example, the private contractor might invest less in secrecy than the government would prefer), because that contractor would design a better system. The revolution in telecommunications brought to light huge benefits of private ownership of phone companies from the perspective of creation and adoption of new technologies (Winston, 1998). In industries where innovation is crucial, the case for government provision is extremely farfetched.

Perhaps as importantly, the adverse effects of cost reductions on non-contractible quality present less of a problem when consumers buy the good and service themselves, and there is enough competition between suppliers that consumers have some choice. Take, for example, the case of health insurance plans or nursing homes, where an important concern is the underprovision of quality by private suppliers once they have signed a customer. In these instances, consumers can switch suppliers if they are dissatisfied with the service. As a result, whether the consumer or the government pays for the service, a seller who reduces cost and quality would find lower demand, or alternatively only demand by the customers who prefer lower quality and cost. It is true that switching may be costly, consumers may be slow to react, and new suppliers may also cut corners. Still, choice between competitive suppliers radically weakens the case for government provision because it weakens the incentives for inefficient cost reduction, while keeping those for efficient cost reduction and innovation strong.

An additional competitive mechanism is reputation-building. If private suppliers want to have customers in the future, they are less likely to cut costs and reduce quality when doing so is inefficient. The reputational effect is evidently of some importance in support of privatization of prisons, where quality innovation is unimportant and prisoner choice is not a realistic possibility, but some private firms work hard to establish a reputation for quality so that they can win additional government contracts (Logan, 1990).

These considerations point to a rather narrow set of circumstances in which government ownership is likely to be superior. These are the situations in which:

1) opportunities for cost reductions that lead to non-contractible deterioration of quality are significant;

2) innovation is relatively unimportant;

3) competition is weak and consumer choice is ineffective; and

4) reputational mechanisms are also weak.
This list gives, in my opinion, a fair sense of how tenuous, in general, is the normative case for government production.

Even this case, however, is still exaggerated. The reason is that market economies have developed a private alternative to profit-maximizing suppliers that attenuates the incentives for quality-reducing cost-reduction, namely not-for-profit provision. Nonprofit firms use the surplus they generate to consume perquisites, to improve the lives of their employees, and even to increase quality when these nonprofits are socially motivated (Weisbrod, 1988). Presumably, the entrepreneurs who run these firms value the perquisites less than the for-profit entrepreneurs value profits, and may in addition care about quality for its own sake. Both of these factors stunt their incentives to make quality-reducing cost cuts. As Glaeser and Shleifer (1998) show, entrepreneurial not-for-profit firms can be more efficient than either the government or the for-profit private suppliers precisely in situations described in the previous paragraph—those where soft incentives are desirable, and competitive and reputational mechanisms do not soften the incentives of private suppliers. Indeed, schools, universities, hospitals, day care centers and other firms that might in principle raise concerns about private provision (whether justifiable or not) often have a not-for-profit status, further undermining the case for government provision.  

The set of activities that are left for the government to perform under these circumstances is very limited, but not empty. As an example, take the case of operating Air Force One, the airplane that carries the President of the United States. Should the operations and servicing of Air Force One be provided by government employees, or by a private contractor? In this example, the argument for government operations is compelling precisely for reasons outlined in this section. First, the security of the president of the United States is important enough that it is a bad idea to have a service provider who might shirk on the quality of personnel, preparedness or service in order to reduce labor costs, or pick shorter and riskier routes to conserve fuel. Soft incentives when it comes to cost containment are essential. Second, ongoing innovation is probably not an important dimension of this particular service. Third, the president cannot readily decide to switch suppliers when his plane is not ready on short notice. Fourth, reputational considerations may not be significant enough to counter the concerns raised above. Finally, the incentives of a not-for-profit operator may still be too strong to contract out this service. While there are of course other examples where government provision is superior, this example gives a sense of the kind of argument that needs to be made.

To conclude this section, note that the case for private ownership made in this paper rests crucially on the importance of incentives to innovate or to reduce costs. In some situations, neither innovation nor cost efficiency is crucial, and contractual incompleteness comes from the government not knowing exactly what it wants, and wishing not to pay so much when it changes its mind (perhaps because raising tax revenue is expensive). In such situations, there is a case for government ownership based

---

3 As Coase (1974) documents, both for-profit and not-for-profit firms operated lighthouses in England, and succeeded in collecting fees from ship-owners, contrary to the economists’ standard view—as expressed, for instance, in Samuelson (1948)—that lighthouses require government construction and operation.
on cost savings. The best examples here are nationalization or heavy regulation in a war, which probably stunt the incentives to innovate, since the returns are largely captured by the government. (Of course, during a war, patriotism may be a separate and powerful source of such incentives.) But a further effect of nationalization or regulation is to allow the government to change its mind about what it wants to be produced, how fast, and by whom, without having to pay a contractor for changing the terms. The alternative of having private contractors submit low bids for government production, in anticipation of charging the government through the nose when it changes its mind, is not efficient when capital markets are imperfect or the government is not credible. This simple argument explains why some government nationalize industries during war, or why they keep such activities as the army or foreign service under state control even in peacetime: it saves money to do so. Indeed, war analogies were prominent in the 1940s writings in favor of socialism and state ownership. It is much less obvious why the government needs to be able to change its mind so often in peacetime without paying for it, and why incentives should be ignored in favor of humoring government unpredictability, especially when the motivating power of patriotism is weakened.

THE POLITICS OF GOVERNMENT OWNERSHIP AND PRODUCTION

So far, I have discussed ownership under the assumption that the government maximizes social welfare. This assumption is too optimistic. Indeed, the principal case for privatization rests on the distortions in the government’s objectives. The discussion of the optimal provision mode is seriously incomplete without paying some attention to the objectives of the government.

Since politicians typically like to remain in power and enjoy the perquisites of their office, a significant element of the goals of any government is to maintain political support. In democracies, such political support usually takes the form of votes, though less democratic governments also need loyalists who can provide manpower to suppress the potential opposition. Governments throughout the world have long directed benefits to their political supporters, whether in the form of jobs at above-market wages or outright transfers. Governments have used their control of state firms and other assets as a means of channeling these benefits, by forcing excess employment at state firms and agencies, creating government projects that transfer wealth to supporters, and so on. In other words, state firms are inefficient not just because their managers have weak incentives to reduce costs, but because inefficiency is the result of the government’s deliberate policy to transfer resources to supporters (Shleifer and Vishny, 1994, 1998; Bennedsen, 1998).

A large body of observation and research is consistent with this view of public production. For three decades after World War II, the British government paid a fortune to keep the country’s coal mines open. It did so not because it was unsure whether the mines were operated efficiently, and not because public managers had weak incentives, but because the miners’ union had the political power to bring down a government. The World Bank (1995) describes a Turkish state-owned coal-mining company that runs annual losses per worker equal to six times the per capita national income; a state-owner power utility in the Philippines that shuts off electricity for seven hours a day in many
parts of the country; a state sugar-milling monopoly in Bangladesh that employs 8,000 unneeded workers while forcing the price of sugar in the country to stay at twice the international level; and a Tanzanian state-owned shoe factory which, even with the World Bank’s help, could not get its production to rise above 4 percent of capacity before eventually shutting down. Consistent with this view of state ownership, privatization of state firms leads to significant improvements in productivity and profitability as employment is reduced and efficiency rises toward private standards (Megginson et al., 1994; La Porta and Lopez-de-Silanes, 1997; Frydman et al., 1998). There is increasing evidence as well that private firms have substantially lower costs and higher productivity in many services that governments have traditionally provided, such as garbage collection (Savas, 1982).

Along the same lines, trade unions around the world are typically the strongest opponents of privatization, precisely because they obtain significant benefits for their members from the government-owned firms in exchange for political support (Lopez-de-Silanes, 1997; Lopez-de-Silanes, Shleifer and Vishny, 1997). The social case for government ownership is the weakest when patronage—the transfer of wealth to constituents through the use of government-owned assets in exchange for political support—is common. At the same time, the prevalence of patronage also explains why government provision is much more widespread than is socially desirable.

Perhaps the most extreme example of using government ownership to pursue political goals is communism. To keep the ruling dictators in office, communist governments of the 20th century came to own entire economies, to plan the allocation of nearly all resources, and to control their populations through monopsony over employment. Such economic strategies proved to be a disaster for the living standards in the communist countries compared to the capitalist countries. Moreover, when economic control did not deliver the communist parties sufficient political security, they killed off their own populations. According to a recent calculation (De Long, 1998), during the 20th century, communist governments killed over 100 million of their own people in peacetime—far more than were killed in all of the century’s wars. The communist tragedy is perhaps the most telling evidence of welfare consequences of government ownership of the means of production, and it should not be forgotten in discussing the niceties of the theoretical aspects of the problem.

In addition to using government ownership to pursue political goals, politicians are also interested in their own income, including bribes. Corruption has two effects on optimal ownership. On the one hand, provision by government employees often puts them in a position to take bribes. In many countries, state tax collectors take bribes to reduce taxpayers’ bills, state policemen take bribes to provide protection, and state doctors take bribes to treat patients. Indeed, some developing countries, including Argentina, Indonesia, Nigeria and Kenya, have contracted parts of customs collections to private Swiss firms with the principal goal of reducing corruption (Low, undated). On the other hand, the process of privatization is itself susceptible to corruption. In exchange for campaign contributions or bribes, politicians may award contracts or sell whole firms to
inefficient providers, overpay these providers, fail to make them accountable for quality, and even fail to enforce those contracts.

Theoretically, then, the effects of corruption on the optimal provision mode are ambiguous: a corrupt government is less able to privatize, regulate or contract in the public interest, but it is also less able to run enterprises in the public interest. In practice, however, the prevalence of corruption typically strengthens the case for private ownership because it is generally easier for reformers in a government to design a relatively corruption-free privatization program, which relies on the effectiveness of a relatively small agency, than to fight corruption inside state firms and agencies (Kaufmann and Siegelbaum, 1997). Thus several Latin American and east European countries, and even Russia, succeeded in designing relatively corruption-free privatization programs despite widespread corruption in their bureaucracies (Boycko, Shleifer and Vishny, 1995). Moreover, once an activity is privatized, the scope of government control and regulation over its delivery generally falls, and so do the opportunities for corruption.

The last point has further implications. The prevalence of corruption is to a large extent endogenous, and depends on the prevalence of government regulations in the economy. In some countries, corrupt politicians create regulations so that they can release firms from them in exchange for bribes (Shleifer and Vishny, 1993). Taking the level of corruption as given in designing the appropriate provision mode is therefore inappropriate. Indeed, the most effective production in some countries may be neither that by the ineffective government nor the contracting to private suppliers by the corrupt politicians, but rather totally unregulated market supply! While such supply may not result in the attainment of all the social goals, it would have the benefit of both significant efficiency and reduced corruption. In an economy with weak institutions, it may be better to have totally private garbage collection than either garbage collection by government employees, or that by the private contractors who got their concessions by bribing officials. It may, for similar reasons, be better to have unregulated private schooling.

The case where provision by a corrupt government remains optimal is one where weak incentives are desirable, market mechanisms are limited, and continuous regulation is necessary but is jeopardized by corruption. In these situations, it may be better to have soft incentives of government employees who pay no attention to quality, than hard incentives of private contractors who bribe government employees to overlook their inattention to quality. It may be plausible to argue for private incarceration in the United States, where elaborate procedures to develop contracts between governments and private suppliers exist, governments have substantial monitoring capacity, courts function in cases of disputes, private suppliers attempt to establish reputations, and corruption is limited. It is quite another matter to advocate such a strategy in a country with weak institutions, where contracting can result in “just take them and bring them back in 10 years” contracts, no protection of prisoner rights by either contracts or courts exists, and contracts that are awarded in return for bribes to government cronies who have no interest in delivering quality service. Such concerns would convince me to oppose prison privatization in Russia, for example.
In summary, political considerations generally not only strengthen the case for privatization, but actually are the crucial reason for it in the first place. Elimination of politically motivated resource allocation has unquestionably been the principal benefit of privatization around the world. Government corruption raises a number of concerns about the design of privatization programs, but in most cases also strengthens the case for private provision. Taking politics into account further reduces the optimal range of government ownership.

THE CASE FOR PRIVATE SCHOOLING

In the United States, one of the most heated recent debates on the optimal mode of provision concerns elementary and secondary education, which is now overwhelmingly provided by public schools (Poterba, 1996). Critics of the current system point to the enormous increases in real public expenditure on schools in the United States in recent decades, which have evidently not been accompanied by corresponding improvements in the quality of education. Hanushek (1996) reports that per student real expenditure on schooling tripled between the mid-1960s and mid-1990s, while standardized test scores, if anything, declined. The opponents of school privatization and choice, in contrast, typically make arguments about quality and social goals. The framework discussed in this paper may be helpful in framing some of the issues in the schooling debate.

For concreteness, I compare two models of provision of education, which, while not capturing every nuance of the existing system and proposed alternatives, do reflect many of the elements of the schemes that are under discussion. Under the current system, schools are controlled by local school boards, which can be thought of as local governments. Schools are paid for through local property taxes and transfers from the state budgets. Schools are therefore generally government-owned and -financed, and the government sets their curriculum and hires teachers. Teachers in public schools are members of a politically active trade union. In general, students have very little choice of schools. Some private schools exist and are financed through private donations and tuition. Parents who send their children to private schools therefore pay private tuition on top of the property and state taxes that go to finance public education.

Under some of the proposed reforms, parents receive vouchers of some fixed value that are paid for, as currently, with local and state taxes. The value of the voucher is initially set to be high enough that it can pay for the currently available public education. Students can use these vouchers at any school of their choice, which can be either public or private. For this scheme to work, it is important that a voucher be accepted in enough schools that meaningful choice can be exercised by the students and their parents. The government can certify the schools at which vouchers can be used, set the basics of the curriculum, and require that students pass certain tests as part of the school recertification process. Private schools can charge more than the value of the voucher if they choose, although the government can also require them to take a certain number of students without any surcharges.

To begin comparing the two systems, note first that from the contracting perspective presented in this paper, the currently existing arrangement is unlikely to be a success, for
at least three reasons. First, because schools are run by government employees, their incentives for cost reduction must be very weak. The explosion of educational costs is not surprising. Second, for the same reason, the incentives for innovation must be very weak. Since ideas of how students should be taught as well as what they should be taught seem to be changing rapidly, the low incentives for innovation are a serious problem. Third, because schools are run by local governments that face a powerful and well-organized labor union, they are likely to put a lot of weight on the welfare of the members of this union as opposed to the consumers of education. Not surprisingly, all these theoretical criticisms of public education in the United States appear to be empirically relevant (Hanushek, 1996; Peltzman, 1993, 1996; Chubb and Moe, 1990).

From the contracting perspective, the proposed alternative of vouchers and school choice has several obviously desirable characteristics. First, private schools have stronger incentives to reduce costs and, more importantly, to innovate, leading in principle to both higher quality and greater efficiency of education. Second, private schools are less likely to cater to the preferences of teachers’ unions. Third, school choice and competition between schools are likely to moderate significantly the incentives of the private providers to undertake quality-reducing cost reductions, since schools that pursue these policies are likely to lose students. Indeed, competition is likely to increase the quality of both private and public schools (Hoxby, 1995). Fourth, reputational considerations are likely to reinforce these benefits of competition. Finally, since the not-for-profit options is the prevalent organizational structure of private schools, their incentives for inefficient cost reduction are further softened. In sum, schooling on the surface appears to be a near-perfect case for the superiority of private provision.

Nonetheless, critics of voucher schemes have produced a number of objections. I will try to list some of these objections, and to evaluate their merits from the contracting perspective.

First, critics argue that there are large social externalities to education, so a private system underprovides this service. As Poterba (1996) points out, there is precious little evidence on the importance of these externalities. Moreover, since vouchers are financed with taxes, it is perfectly possible for the government to maintain higher educational spending than what consumers would do if left to their own devices. Broad externalities, therefore, are not a compelling objection, even if they exist.

Second, critics argue that there are large within-school external benefits of heterogeneity, whereby poor students benefit from the presence of students from wealthier households, and perhaps vice versa. They claim that a voucher scheme would therefore lead to socially suboptimal segregation. As Hoxby (1996) clearly shows, there is a significant amount of segregation of students in the United States already, driven by geographical mobility of households and their choice of neighborhoods. It is quite possible, therefore, that a voucher scheme would reduce segregation if students from poorer households have the choice of going anywhere they want, and schools are required to take at least some students without charging more than the voucher as a condition for the government redeeming their vouchers. One extreme mechanism of combating segregation is
Milwaukee’s random allocation of voucher-holding students to schools. Indeed, vouchers may be a ticket to good education for many students (Moe, 1995; Rouse, 1998).

Third, critics are concerned that some private schools will have curriculums that do not fully reflect the social goals of education. Many people have indeed argued that inculcating students with patriotism and ideology is the crucial reason for public education in the first place (La Chalotais, 1763, quoted in Barker, 1944, p. 80; Lott, 1990; Kremer and Sarychev, 1998). Prussia and France, for example, developed state education in part for ideological reasons—the former to foster religion, the latter to counter its influence (Barker, 1944). In the United States, there has been a fear that, if some private schools are captured by groups with peculiar religious or cultural views, ideological goals of public education will be undermined (Kremer and Sarychev, 1998).

While such concerns may have been important at the time of creation of public education, when contracting opportunities were limited, they are not as compelling today. To begin, as someone who went through Soviet schools, I am underwhelmed by the alleged benefits of indoctrination. But even someone who believes in large social benefits of teaching state ideology must recognize that these problems can be largely addressed contractually, by requiring that particular subjects be taught and others not taught, stipulating the basic curriculum, and testing students as a condition for school eligibility for a voucher program. It does not appear that these terms are inherently non-contractible, although they might have been in the 19th century. Thus I am not sure why teaching evolution, and restricting the teaching of creationism, is inherently non-contractible: if anything, private schools might be more concerned about violating their agreements with the government if they face significant financial penalties than the public schools in communities that want creationism taught. And even if some teaching of “undesirable” doctrines under private schooling is unavoidable, the social costs of such teaching should be compared to the potential benefits of improved education resulting from improved incentives.

Fourth, critics worry that private schools will attempt to exclude certain students because they are troublemakers, have learning disabilities, or may otherwise be expensive to educate. This too does not seem to be a valid objection, since presumably the government currently pays more either to the existing public schools or to special (private or public) programs to help these students. Such programs can always continue. The key point is that, as long as entry of schools is allowed, and vouchers are generous enough that schools want to attract students, the first-order effect is to make access to good education easier rather than harder for everyone.

Finally, some of the critics argue that parents are too uninformed to exercise meaningful choice (Arrow, 1996). Although this may be true in some cases, it is important to remember that the kind of knowledge necessary for school choice is not that of the specifics of the courses and academic curricula. Most people would probably choose schools by checking with their neighbors, and most people are capable of such an inquiry. The children of parents who do not check will probably find themselves in no worse a predicament than they face today. After all, competition is likely to improve the overall quality of public schools if they want to stay in business (Hoxby, 1995).
In summary, at least the theoretical case of school choice and vouchers as an alternative to the current system in the United States seems overwhelming. The objections to this case appear to be insufficiently focused on the true possibilities of contracting and regulation, and on the enormous benefits of private ownership and competition for both cost efficiency and innovation.

CONCLUSION

Private ownership should generally be preferred to public ownership when the incentives to innovate and to contain costs must be strong. In essence, this is the case for capitalism over socialism, explaining the “dynamic vitality” of free enterprise. The great economists of the 1930s and 1940s failed to see the dangers of socialism in part because they focused on the role of prices under socialism and capitalism, and ignored the enormous importance of ownership as the source of capitalist incentives to innovate. Moreover, many of the concerns that private firms fail to address “social goals” can be addressed through government contracting and regulation, without resort to government ownership. The case for private provision only becomes stronger when competition between suppliers, reputational mechanisms, and the possibility of provision by not-for-profit firms are brought into play. Last but not least, the pursuit by government officials of political goals and personal income, as opposed to social welfare, further strengthens the case for private ownership, as the dismal record of state enterprises around the world and the tragedy of communism illustrate all too well.

The benefits of private delivery—regulated or not—of many goods and services are only beginning to be realized. Health, education, some incarceration, some military and police activities, and some of what now is presumed to be “social” insurance like Social Security, can probably be provided more cheaply and attractively by private firms. It is plausible that 50 years from now, today’s support for public provision of these services will appear as dirigiste as the 1940s arguments for state ownership of industry appear now. A good government that wants to further “social goals” would rarely own producers to meet its objectives.


ABSTRACT

This Note applies agency costs theory to explain charter schools’ use of for-profit and nonprofit forms, and to suggest ways to make charter school regulation more sensitive to the differences between these forms. Borrowing from Henry Hansmann’s “contract failure” theory of nonprofits and recent data on the makeup of the charter school market, I argue that nonprofit forms dominate because they minimize the unusually high agency costs that characterize interactions between charter
operators and the parents, regulators, and donors who influence them. For-profit schools survive only when the economies of scale they capture through superior capital-raising offset their higher agency costs. I also compare nonprofits’ and for-profits’ abilities to achieve some of charter school policy’s more complex goals. These include resource attraction, localized governance, and output-based accountability. I conclude by arguing for changes in regulation to control for-profits more tightly and to reflect more accurately nonprofits’ and for-profits’ relative strengths.

INTRODUCTION

Charter schools are becoming an increasingly important part of America’s primary and secondary education system. Since 1991, forty states, the District of Columbia, and Puerto Rico have authorized charter schools, and more than 3000 such schools have opened. Congress has recently devoted considerable time and attention to charter schools, authorizing major grant programs to assist charter schools with startup costs and facility acquisition, and allowing failing conventional public schools to restructure themselves as charter schools under the No Child Left Behind Act.

As charter schools increase in importance, so does the debate about the laws that govern them. In this Note, I examine one aspect of that debate: whether to organize charter schools as nonprofit or for-profit entities. The role of profit has become a contentious issue in charter school policy. For-profit schools have drawn criticism for cutting quality in the interests of shareholders and overlooking the public good aspects of education. They have also drawn praise, however, for the innovation, efficiency, and market-style discipline they promise to bring to education.

I explore charter schools’ use of the nonprofit and for-profit forms from the perspective of agency costs theory. Though the debate about charter schools is so broad that strong claims about its content are dangerous, voices well informed by agency costs theory appear to have been largely absent. This Note seeks to fill that gap.

Agency costs theory illuminates the debate about charter school organizational form in a number of ways. First, agency costs theory explains the relative dominance of nonprofits in the charter school market. As Henry Hansmann has argued, the nonprofit form is useful when principals have an unusually difficult time monitoring or enforcing contracts with their agents. By prohibiting nonprofits from distributing earnings to shareholders, state laws allow nonprofits to pledge with a high degree of credibility that they won’t cheat their patrons in ways the patrons can’t perceive or control. Since nonprofits don’t stand to gain anything from cheating, they have little reason to do it. Hansmann’s model fits the charter school market, because the various parties that control charter schools’ success—parents, government agencies, and donors—all have a great deal of difficulty perceiving and controlling abuse and cheating by charter schools. In other words, monitoring costs are high and the nonprofit form reduces them. A significant minority of for-profit charter schools exists in spite of these high monitoring costs, however, because for-profit schools’ superior access to capital markets allows them to operate on a large
scale. For-profit schools exist when they can achieve economies of scale that outweigh agency-cost inefficiencies.

Second, agency costs theory suggests that nonprofits may be better than for-profits at furthering some of charter school policy’s more complicated goals. Nonprofit schools attract charitable donations and improve local teachers’ and parents’ control more effectively than for-profits do. This conclusion must not be overstated; for-profit schools may be slightly more responsive to the pressures of output-based accountability than nonprofits are. But there is little doubt that organizational form has important consequences for charter schools’ abilities to actualize the goals of charter policy.

Finally, agency costs theory suggests that regulators may need to control for-profit schools more aggressively than nonprofit schools. The rhetoric of the charter school movement has emphasized charter-school-specific forms of accountability, such as parental monitoring and periodic output-based review by government agencies. But agency costs theory suggests that the nonprofit form may be just as important in regulating most charter schools as these more direct forms of accountability are. For-profit schools not subject to the nonprofit form’s constraints are therefore likely to expose the gaps and cracks in charter-school-specific regulation. In this Note, I suggest a handful of regulatory changes to meet the challenges for-profit schools pose. These include curtailing “hybrid” for-profit-nonprofit management arrangements, limiting for-profit schools’ ability to cut quality in imperceptible ways, and strengthening existing governmental monitoring and information-gathering systems.

I begin in Part I by explaining the charter school concept and the organizational forms charter schools use. These forms tend to be complex and cannot be placed into “nonprofit” and “for-profit” categories as neatly as forms in most industries. In Part II, I explain nonprofit schools’ relative dominance. In Part III, I explain why nonprofits’ dominance is not total, and why there are still a significant number of for-profit charter schools. I move in Part IV to examining how for-profit and nonprofit schools differ in achieving some of the more complicated goals of charter school policy. In Part V I then propose changes in regulation to meet the challenges posed by for-profit schools.

I. CHARTER SCHOOL STRUCTURE

“Charter schools” are notoriously hard to define. Charter laws differ widely from state to state, and charter schools exist alongside other school choice regimes, some of which go by different names but are difficult to distinguish from charter schools. The concepts of “for-profit” and “nonprofit” are also complicated in the charter school context. Mostly to avoid restrictions on for-profit entities’ abilities to hold charters directly, charter schools have produced unusual fusions of for-profit and nonprofit legal forms that defy easy categorization. To clarify these murky concepts, in Section A I define “charter schools” and distinguish them from conventional public and private schools. In Section B, I draw distinctions among charter schools, and separate these schools into for-profit and nonprofit groups.
A. The Charter School Concept

Charter schools combine elements of conventional public and private schools. The key characteristic of a charter school is that it combines public funding with private management. Unlike conventional private schools, charter schools do not charge tuition and receive all of their funding from state and local governments, school districts, and private charitable donations. Unlike conventional public schools, however, charter school teachers and managers usually are not government employees. State laws place few restrictions on who can start a charter school, and the charter school definition the U.S. Department of Education uses for administrative purposes allows parents, teachers, school administrators, and any “members of the local community” to operate a charter school.

Enrollment in a charter school is also usually optional. Unlike conventional public schools, charter schools rarely have geographic boundaries inside which attendance is compulsory. Students can leave charter schools for other charters, conventional private schools, or conventional public schools. If a school has more students wishing to attend than seats, it must admit students by lottery.

Prior to opening, a charter school must receive a “charter” from a statutorily authorized agency. Under the federal definition, any “public entity” authorized under state law and approved by the Secretary of Education can authorize a charter school. In most states, charters may be granted by local districts, state departments of education, or universities. The chartering process and the level of scrutiny applied to charter applications vary widely among states.

A related point is that chartering agencies rarely solicit applications. Unlike most publicly funded, privately operated enterprises, charter schools typically apply for public funding on their own initiative, operate largely according to their own terms, and rarely have to endure a competitive bidding process. When cities hire private trash collectors, for example, they usually begin by defining the areas to be served and seeking competitive bids. In most cases, after signing a contract, the city stops collecting the trash and lets the winner of the bidding process take over. In contrast, school-chartering agencies tend to wait passively for charter operators to come to them, and the goals of charter schools may have little or no connection to the strategic plans of the districts in which they operate. Charter applicants usually determine their own “educational objectives” and basic curricula; these objectives and curricula are merely “agreed to” by chartering agencies. And charter authorizers usually grant charters to worthy applicants even when existing conventional public schools have the capacity to educate all students in an area. This often means that conventional schools’ costs remain the same even after charter schools appear in their areas and begin drawing away students and the funding attached to them. Sometimes charter schools follow a more traditional privatization model. Under the No Child Left Behind Act, for example, local districts, on their own initiative, can dissolve failing conventional public schools and convert them to or replace them with charter schools.
Under the U.S. Department of Education definition, charter schools are “exempt from significant State or local rules that inhibit the flexible operation and management of public schools.” Granting charter schools greater freedom ideally enhances their ability to innovate and operate efficiently. Unlike conventional public schools, which the government monitors by dictating inputs—e.g., whom to hire, what to teach, and where to teach it—charter schools in theory are primarily regulated by the monitoring of their output as measured by student achievement. They develop their own educational objectives and the means for measuring them, and government agencies do not interfere with these decisions as long as charter schools demonstrate success. When charter schools fail, chartering agencies or state boards of education theoretically respond by closing the schools rather than by dictating inputs.

The “output-based regulation” ideal has frequently faltered in practice. Charters have often failed to establish meaningful and measurable goals, and charter schools’ freedom is limited by significant input-based regulation. Many (perhaps most) charter schools must comply with state-level regulation of curriculum and teacher qualifications. And exemptions from state laws and regulations appear to have been rare. Charter schools must also be nonsectarian, and must comply with antidiscrimination, auditing, safety and health, and other applicable state and federal laws. Nevertheless, charter schools have often found ways to carve out room for unique curricula and governance structures even in the face of heavy regulation.

B. Distinguishing For-Profit and Nonprofit Charter Schools

Charter school operators have discovered a variety of innovative ways to combine for-profit and nonprofit forms. These forms occupy a continuum, with schools owned and operated by for-profit entities on one end and schools owned and operated by nonprofit entities on the other. Although the charter school literature has often failed to distinguish among these forms, a thorough understanding of them is essential for my purposes. The place a school occupies on the continuum will determine how accurately economic models developed to describe more pure for-profit and nonprofit firms apply.

On the nonprofit end of the continuum are charter schools organized under state laws as nonprofit corporations that qualify for tax exemptions under I.R.C. § 501(c)(3). In a pure nonprofit charter school, the nonprofit entity that holds the school’s charter manages all strategic and day-to-day operations and directly employs all the teachers, administrators, and staff. Also near this end of the continuum are nonprofit charter-holding organizations that hire nonprofit management entities to operate their schools. In these schools, one nonprofit entity may receive the charter from a chartering agency, and then hire a separate organization to actually manage the school. On the for-profit end of the continuum are firms organized as for-profit business entities under state law that both hold charters and manage their schools’ operations. Charter schools on the extreme for-profit end of the continuum are rare.

The middle of the continuum is inhabited by what I will call “hybrid schools.” In a hybrid school, a nonprofit entity receives and holds the school’s charter, and contracts with a for-profit firm for management services. Sometimes these arrangements make genuine
economic sense, with nonprofit charter-holding entities and for-profit management firms each playing their roles more efficiently than the other could. Often, though, these arrangements owe their existence to state laws that prohibit for-profit entities from holding charters directly. Since they can’t hold charters themselves, for-profit entities find or create nonprofits capable of doing it for them. Nationally, approximately fourteen to nineteen percent of nonprofit charter schools contract with for-profit management firms for at least some services.

The place a particular hybrid school occupies on the for-profit-nonprofit continuum depends on the kinds of services the for-profit management firm provides. Some hybrid schools hire for-profit firms only for limited logistical services or the outlines of a curriculum, and thus fall close to the nonprofit end of the continuum. Most hybrid schools, however, grant for-profit management companies more substantial roles. The U.S. Department of Education found that seventy-one percent of hybrid schools hired for-profit firms to “manag[e] the overall operation or administration of [the] school.” Sixty-four percent had the for-profit firms direct curriculum and instruction, and sixty percent had the firms hire staff. Sixty-four percent of the nonprofit charter-holding entities in hybrid schools also received seed or startup funds from their for-profit managers.

More detailed observation supports the story these national statistics tell of deep involvement by for-profit firms. A report issued by Western Michigan University found that for-profit management companies in Michigan—a state with an unusually high percentage of for-profit schools—often own charter schools’ buildings, equipment, and supplies; nominate and cultivate support for board members of the nonprofit entities that apply for charters; and contribute startup capital. In fact, some management companies in Michigan refuse to contract for anything other than “full service” agreements that grant them total authority over the schools. The existence of a number of for-profit management companies serving only one school suggests that the line between nonprofit charter-holding entities and their for-profit management companies is thin. A for-profit management firm serving only one school would have no economies of scale to justify independence from its client, absent legal restrictions on direct ownership by the management firm. Anecdotal evidence suggests that many nonprofit entities are merely fronts for for-profit firms.

The place a hybrid school occupies on the continuum may also be influenced by the way its management firm is compensated. Two basic arrangements are possible. A management company may receive a fixed fee (calculated per student, per school, or by some other method), and the nonprofit hiring organization may pay all of the school’s operating expenses. Under this kind of arrangement, a for-profit firm might have relatively little to gain from cutting costs or quality, except to the extent that overspending might make the nonprofit charter-holding organization insolvent and eliminate the possibility of a continuing relationship. A second possible arrangement may grant a management company a fee and require the company to pay its operating expenses out of that fee. In these circumstances, the for-profit management company has much to gain by cutting costs, since it receives all of whatever surplus remains after expenses are paid. In this kind of compensation arrangement, the for-profit manager may
be indistinguishable from an owner who receives nearly all or most of the business’ revenues and bears the requisite level of liability and costs.

Evidence of how for-profit management firms are actually compensated is hard to find. Its last public filing before it went private suggests that Edison Schools, the nation’s largest for-profit education management company, typically receives a large fee and pays its own operating expenses out of that fee. But it is unclear if Edison’s experience is typical. Only a handful of for-profit education management organizations are public companies subject to filing requirements under the securities laws, so little is known about their operations. These companies naturally hesitate to share the details of their management contracts with anyone other than potential investors.

The place that hybrid schools occupy on the for-profit-nonprofit continuum affects how closely economic models developed to describe for-profit firms can be applied to hybrid schools. Hybrid schools that hire for-profit managers only for limited purposes may closely emulate pure nonprofit schools in their economic behavior. On the other hand, hybrid schools in which the for-profit management firms’ control and incentive are powerful enough may behave similarly to schools fully owned and operated by for-profit entities. For example, if a for-profit management firm controls all of the day-to-day and strategic operations of a hybrid school, and the firm’s compensation turns in some way on its ability to minimize costs and maximize revenue, a hybrid school has nearly the same incentive and capacity to lower costs and increase revenue as a pure for-profit school. In such situations, the management company internalizes most of the upside and downside of owning the firm. It also enjoys most of the control, as the nonprofit charter-holding firm is unlikely to exercise meaningful authority. Even if a non-profit board is more than a stand-in created by the for-profit management firm to obtain the charter, the for-profit management firm’s control may be so complete that firing it would be akin to closing the school.

In keeping with this understanding of hybrid for-profit firms, in the rest of this Note I use the term “for-profit” to refer both to pure for-profit schools and to hybrid schools in which the for-profit management firms’ involvement is so deep that the schools function, in economic terms, like for-profit firms. Similarly, I count as “nonprofit schools” hybrid schools in which the for-profit management firms’ involvement is too small to distinguish the schools from pure nonprofit schools.

... [According to an Arizona State University study,] nonprofits have consistently dominated the charter school market since 1998-1999, the first school-year for which comparative figures are available. Although for-profits have never come close to competing with nonprofits for market leadership, [accounting for 10 to 14 percent of total charter schools,] they have nonetheless maintained a presence that must be accounted for.

There are two limitations in these data worth noting. First, the ASU report included as for-profits only schools that were completely managed by for-profit entities. That is, the ASU report did not consider a hybrid school to be under for-profit management if the for-profit firm had anything less than total control over the school’s day-to-day and strategic operations. This may bias the number of for-profit schools slightly downward, if we think
a hybrid school may act like a for-profit firm even when it’s under less than total control by a for-profit firm. Second, the report likely failed to include some hybrid schools managed by small for-profit firms. For-profit companies with only one or two schools under management tend to keep a low profile, which makes them difficult to find and identify in a nationwide study. This biases the percentage of charter schools downward, since it means there are hybrid charter schools out there that the ASU study did not find. The ASU study’s disproportionate failure to find small for-profit operators is primarily important for my conclusions about for-profit firms’ tendency to operate at large scales.

II. EXPLAINING NONPROFIT DOMINANCE

In this Part, I argue that nonprofits dominate the charter school market because they control agency costs more efficiently than for-profits do. In Section A, I explain Henry Hansmann’s “contract failure” theory of nonprofit organizations, focusing on the theory’s prediction that industries characterized by high monitoring and agency costs tend to be dominated by nonprofits. In Section B, I argue that this theory fits the charter school market because monitoring and agency costs are high for many of the constituencies that influence charter schools’ survival. In Section C, I analyze in detail how the constraints of the nonprofit form reduce agency and monitoring costs in charter schools. In Section D, I consider some alternative theories to explain nonprofit entities’ dominance of the charter school market.

A. Contract Failure Theory

Although economists have yet to settle on a paradigmatic general theory of nonprofit firms, the last two and a half decades have seen Henry Hansmann’s contract failure theory begin to prevail. The theory emphasizes nonprofits’ ability to reduce the inefficiencies stemming from high monitoring costs. Hansmann explains the theory succinctly:

[N]onprofit firms serve particularly well in situations characterized by . . . “contract failure”—that is, situations in which, owing either to the nature of the service in question or to the circumstances under which it is produced and consumed, ordinary contractual devices in themselves do not provide consumers with adequate means for policing the performance of producers. In such situations, the nonprofit form offers consumers the protection of another, broader “contract”—namely, the organization’s commitment, through its nonprofit charter, to devote all of its income to the services it was formed to provide.

Contract failure may occur for a variety of reasons. Consumers may be unable to compare different producers, make a clear bargain about the nature of the goods and the price to be paid, evaluate a product, or enforce a bargain. In situations characterized by contract failure, producers can cut quality in ways that consumers either can’t perceive or to which they can’t respond. For-profit firms are particularly dangerous to consumers in these situations because the owners of these firms benefit from whatever surplus the firms gain from their customers’ losses.
Consumers respond to this problem by turning to nonprofit firms who act somewhat like fiduciaries on their behalf. Nonprofit firms can serve this function because state law prohibits them from distributing profits to shareholders. These firms must plow their earnings back into their businesses and devote themselves to serving patrons. This “nondistribution constraint” assures patrons that the firm has little incentive to take advantage of them by reducing quality. This is important for patrons who cannot verify directly that the firm is treating them fairly.

Hansmann offers a number of examples of nonprofit organizations that fit this mold, such as CARE and the American Red Cross, which receive money from donors to subsidize demand among the poor for necessities like food and clothing. Hansmann argues that although it may seem odd, for-profit firms could conceivably provide this same service: for example, they could provide X pounds of rice to a person in Africa in exchange for Y dollars from a donor in the United States. For-profit firms rarely provide this kind of service, however, because monitoring costs are too high; a donor in the United States has almost no way of knowing how well a firm is subsidizing demand in Africa. In the absence of mechanisms for directly verifying a firm’s performance, the donor does the next best thing: She gives her money to a nonprofit firm that, because of the nondistribution constraint, has little incentive to cut quality in unverifiable ways.

B. High Monitoring Costs in Charter Schools

Hansmann’s theory can explain the dominance of nonprofit firms in the charter school market. Several interest groups influence a charter school’s success and failure, including parents, government agencies, and donors. Each of these groups has goals they want the charter school to meet (generally centering on academic achievement), and each faces significant monitoring costs that prevent them from assessing and enforcing the school’s attainment of those goals. These groups face common monitoring problems, including perceiving and measuring students’ achievement and enforcing the threat of accountability. Hansmann’s theory explains nonprofit dominance as a product of these groups’ efforts to reduce monitoring costs. These groups prefer nonprofits because nonprofits offer assurances that managers will not cut costs and quality in imperceptible ways. In the following three Subsections, I explore the monitoring costs each of these groups encounters and explain how these groups respond to monitoring costs by encouraging nonprofit dominance.

1. Monitoring by Parents

Early charter theory placed a great deal of emphasis on parental monitoring. Early charter theorists argued that schools would live or die based on their ability to gain parents’ trust. In fact, parents have had great difficulty monitoring charter schools and holding them accountable. A parent’s goal is relatively simple: She wants the best education for her child. But parents face two obstacles in monitoring schools’ achievement of this basic goal: (1) measuring and assessing output and (2) enforcing accountability once they become dissatisfied.
The first set of problems is rooted in parents’ distance from their children’s education. Parents do not sit in classrooms and receive instruction; their children do. Parents, therefore, must resort to proxy measures of schools’ output, such as tests and conversations with students and teachers. Each of these has severe limitations. The debate about testing is large and complex, and it need not be rehearsed again here. It is sufficient to point out that tests do not seek to measure many relevant character traits and values that education aims to instill in children, such as self-discipline and cooperation. Tests are especially problematic in assessing idiosyncratic charter schools such as Montessori schools, which make character cultivation a central aim. Additionally, test data may not be easily accessible for parents making decisions. While parents may be able to obtain their own children’s test scores, aggregated test data is often hard to acquire. In fact, many states have had severe problems collecting and disseminating information to parents.

Of course, parents can also talk directly to their children and to teachers, but these softer measures have limits as well. Teachers at for-profit schools, for example, may be unreliable sources of information because they feel pressure to speak optimistically to keep students from leaving. Children may also be unreliable: They can be quizzed and questioned, but they often lack the capacity to judge their educational experiences and to articulate those judgments to parents. Finally, children’s goals may conflict with their parents’. A high school student doing poorly at a charter school may prefer to stay at the school because her friends are there, while a parent may prefer her to leave.

A larger problem that plagues almost any method of measuring a school’s output is the risk that a child’s performance depends on factors beyond a school’s control or influence. When a teenager loses interest in school, is it because of the school or because of the teenager? Large-scale studies can sometimes make judgments within reasonable confidence intervals about a school’s general effect on students, but the uncertainty is much greater when assessing individual students. Even if a school does well on measures of aggregate student performance, parents have little insight into whether their own children are outliers and could do better elsewhere.

The second major monitoring obstacle is that even when parents can gather and comprehend information, they may be in a poor position to enforce accountability. Children become attached to teachers, friends, and routines, and even if a school is performing worse than other available schools, the psychological and emotional costs of removing a child may well exceed the uncertain gains from putting her in a better school. Additionally, parents may be vulnerable to a sunk-cost bias, irrationally keeping their children in underperforming schools because they do not want to confront the failures of their own judgment.

Parents may also have few alternatives. Most school choice plans are small, and parents thus face a choice between one or perhaps a handful of conventional public schools and one charter school. For parents who prefer idiosyncratic charter schools, such as Montessori schools or schools with particular subject emphases, the choices are particularly limited.
Hansmann’s contract failure theory suggests that parents may respond to all of these monitoring problems by preferring nonprofit schools. Though probably unfamiliar with the nondistribution constraint, parents may intuitively recognize their own vulnerability to being cheated and may understand that nonprofit schools not subject to profit motives and shareholder pressure are less likely to cheat them than for-profit schools. Parents’ preferences, in turn, influence the shape of the charter school market. If parents don’t trust a school, they can refuse to send their children there. Since funding is tied to enrollment, schools without students go out of business. And since parents have reason to prefer nonprofits, it is not surprising that nonprofits dominate the charter school market.

2. Monitoring by Governments

Government agencies also play an important role in determining charter schools’ success. Like parents, government agencies have great difficulty monitoring charter schools. Hansmann’s theory therefore suggests that agencies use their influence to encourage nonprofit charters more than for-profit charters.

Government agencies monitor charter schools for their achievement of several goals. First, governments want children to learn the material they’re taught. Often this interest overlaps with parents’ interests. Occasionally, however, it may bring governments and parents into conflict when parents are unable or unwilling to seek their children’s best interests. Governments also have public-good-related goals. These generally do not overlap with parents’ interests and may even conflict with them. In a democracy, education serves public ends by molding students into good citizens capable of fulfilling their civic responsibilities. Students’ fulfillment of these duties benefits all members of society. Since good citizens do not fully internalize all of the benefits their civic-mindedness brings to society, education is in some senses a public good. Public-good-related interests may bring governments into conflict with parents who prefer bizarre anti-social curricula, or (more likely) curricula that emphasize specialized subject areas at the expense of subjects that might transform children into better citizens. Governments also have distributional goals, which motivate them to ensure certain norms of equal access and compliance with disability and nondiscrimination laws.

Governments, however, have their own limitations. First, government agencies face many of the same output-assessment and accountability-enforcement problems that parents do. Second, goal-definition may be more difficult for government agencies than for parents, because agencies must reconcile competing ideas of educational success. Third, information-gathering problems are in many ways more complex for governments, which are far removed from individual children and must discern broad trends and make large-scale policy judgments. In making these judgments, governments must rely more heavily on tests than parents do. Finally, agencies face obstacles to using the limited regulatory mechanisms that charter school statutes provide to enforce accountability. Although charter school theorists emphasize the threat of government-mandated closure, such closures have been rare. Charter schools often become deeply politically entrenched and even underperforming schools tend to gather support. A group of researchers who participated in a UCLA-sponsored study of California charter schools has argued that “holding charter schools accountable [i]s as much a political process as it [i]s an
administrative matter.” Charter schools have fought at least as vigorously in political and bureaucratic settings as in the market for students.

Government monitoring has also been hampered by a wide range of failings that, though not inherent in charter theory, arise frequently in practice. For example, charter schools and government agencies are often uncertain about the scope of the government’s monitoring authority. The most recent study of the Public Charter School Program found that “charter school legislation in the states has provided virtually no guidance on how authorizers should approach accountability processes.” One charter operator in California and his district superintendent disagreed even about whether the charter school had to meet the district’s achievement standards or could set its own. In Arizona, which has long been a leading charter school state, researchers observed similar uncertainty. Such vagueness in a monitoring agency’s mandate can profoundly cripple the agency in the politicized charter environment.

Monitoring agencies also frequently lack adequate resources. In fact, most monitoring bodies have no staff specifically devoted to charter school issues at all. The symptoms of this problem have included limited communication between agencies, poor information gathering, and, in some states, the failure of any agency to take responsibility for assessing charter school performance. What few resources agencies devote to charter schools may have limited effect because charter schools often fail to set clear and measurable goals for themselves.

The challenges government agencies face in monitoring charter schools are evident in agencies’ emphasis on indirect forms of monitoring and regulation. Agencies tend to hold charter schools accountable primarily for failures to comply with financial regulations, nondiscrimination regulations, and other definite but narrow rules, rather than for shortcomings in academic performance. Additionally, many charter schools receive no more autonomy from input-based regulation than conventional public schools do, presumably because many regulators do not trust their ability to control charters through output-based monitoring. This reality contrasts with charter theory and statutory program design. Most charter authorizers and monitoring agencies claim to monitor academic performance. Indeed, output-based accountability is a central theme in the charter school movement. But academic performance is so difficult to measure and so open to controversy that poor academic performance has resulted in few school closings.

Hansmann’s contract failure theory predicts that in response to these monitoring difficulties government agencies will prefer dealing with nonprofit charter schools. Though direct evidence of agencies’ preference for nonprofits is hard to find, agencies may have encouraged nonprofits in a number of ways. Agencies determine which schools may open and how long they may operate. In practice, the scrutiny chartering agencies apply to new charter applications may vary widely. But when such scrutiny is meaningful, chartering agencies may be more likely to approve nonprofit schools than for-profit schools because the agencies may recognize the deficiencies in their own ability to monitor. Additionally, to reduce their monitoring costs, agencies may design and police nonstatutory regulations in ways that are unfriendly to for-profit charter schools.
3. Monitoring by Donors

Unlike conventional public schools, some charter schools attract substantial donations. Because donors’ money plays an important role in determining which schools succeed and which fail, donors’ tendency to choose nonprofits may also help explain the shape of the charter school market. To understand the ways in which donors incur monitoring costs, it is important to understand the forms donations take. Charter donors make their contributions in a wide variety of ways. First, they contribute cash. It is difficult to pin down exactly how much money private donors have given charter schools, but a recent U.S. Department of Education study found that the great majority of states reported private cash donations as sources of startup funds for charter schools. Anecdotal evidence also suggests that donations have been important to many charter schools. This evidence is bolstered by the fact that many states, either deliberately or inadvertently, give charter schools less money than conventional public schools, making donations important.

Second, donors contribute volunteer labor. Many researchers have found that charter schools generally get more out of parents and volunteers than traditional schools do. This labor may not be as useful as skilled paid labor, and it may inefficiently distort charter schools’ choices away from cash-and-capital-intensive instructional methods (such as computers) to labor-intensive methods (such as supervised reading and games). But volunteerism can come in the form of valuable skilled labor as well. Amistad Academy, a highly successful charter school in New Haven, Connecticut, for example, has on its board several prominent community members who volunteer their fundraising, finance, and management expertise.

Finally, nonprofit charter schools often attract talented and energetic managers and teachers who forego higher salaries in private schools or other industries to pursue their passion for education. It may seem strange to think of these skilled workers as “donating” their labor, but the economic effect is similar to volunteerism. The enthusiasm and skill of undercompensated workers may well be the most significant factors driving charter school success. Amistad Academy, for example, was founded and operated by a group of prominent community members led by Dacia Toll, a Yale Law School graduate and Rhodes Scholar with substantial management expertise. Toll says that were it not for the opportunity to become involved in a charter school on her own terms, she probably would not have entered public education. Teachers appear to derive satisfaction from the perception that charter schools are more flexible, collegial, and serious about success. As a result, these teachers may work harder and may stay in publicly funded education longer than they otherwise would.

Donors face many of the same monitoring problems that agencies and parents do. Because children’s achievement is so difficult to measure, and because the risk that children’s achievement is influenced by forces outside of the schools’ control is great, donors may hesitate to give their money and time to for-profits. Donors’ ability to monitor may depend to some extent on the form the donations take. Perhaps some large cash donors, important volunteers, and underpaid employees are in better positions to monitor charter schools than parents and agencies are. They may be able to gain
something like insider status, sitting on charters’ boards or participating in their management and decisionmaking.

Even if some donors have some ability to monitor and enforce accountability, there are enough schools vying for these donors’ money that even small differences between for-profit and nonprofit schools in the difficulty of monitoring may cause donors to offer their money and services to nonprofits. If that is so, nonprofit schools gain a competitive edge by having access to donated funds and labor. It is difficult to say exactly how strongly donors have influenced the makeup of the charter market, for the same reasons it is difficult to say how much donors contribute. Contract failure theory, however, strongly suggests that donors will favor nonprofit charter schools.

C. Enforcement of the Nondistribution Constraint in Charter Schools

We now have some sense of the problems parents, agencies, and donors face in monitoring charter schools. We also know that Hansmann’s contract failure theory suggests that these constituencies will use their influence to encourage the growth of nonprofits as a solution to these monitoring problems. In this Section, I turn to the question of how exactly nonprofits solve monitoring problems. The nondistribution constraint in nonprofit charters provides parents, governments, and donors a guarantee that the schools will not cut quality in ways these constituencies cannot perceive to increase shareholders’ profits. In this regard charter schools are a typical instance of the general contract failure story.

Ordinarily, the nondistribution constraint is enforced by state attorneys general, the IRS, or, in a small number of states, patron lawsuits. This kind of enforcement is as plausible in the charter school industry as in any other. Nonprofit charter schools may be subject to closer policing on the nondistribution constraint than nonprofits in other industries, though, because charter schools are regulated heavily in other aspects of their businesses. The nondistribution constraint may sometimes be enforced by state boards of education or other charter regulatory agencies. Although charter regulators have many failings, they appear to police charter school finances relatively well. These efforts may indirectly (and inadvertently) have an effect similar to enforcement of the nondistribution constraint. In the process of auditing charter schools and scrutinizing the way they use government funds, charter regulators may stumble across violations of the nondistribution constraint. Many charter closures have resulted from discoveries of the deliberate mismanagement and indirect profit distributions that typify violations of the nondistribution constraint in more conventional industries. Sometimes charter regulators recognize this kind of dubious conduct as a violation of state nonprofit corporation laws and may enforce those laws directly or report violations to state attorneys general. It is also possible that charter regulators process violations of the nondistribution constraint as violations of charterschool-specific financial regulations or regulations requiring charter schools to comply with all relevant state laws. They may, for example, require particular funds to be spent in particular ways, or may place tight limits on the way schools use public funds. In any event, the effect is the same.
Besides inviting direct and indirect enforcement of the nondistribution constraint, the nonprofit form may also constrain charter schools in extralegal ways. Hansmann has argued that the social norms associated with nonprofit enterprise may be more effective than the threat of legal sanctions in policing the nondistribution constraint. A significant empirical literature has demonstrated that workers at nonprofit firms are willing to accept lower pay than their similarly qualified peers at for-profit firms in exchange for the extra satisfaction that comes from working at nonprofit firms. Since nonprofit charter schools are often filled at least partly with teachers who believe in the schools’ missions, nonprofit charters may be less likely to take advantage of parents’ and agencies’ inability to monitor than for-profit charters are.

D. Alternative Explanations for Nonprofit Dominance

Contract failure is not the only way to understand nonprofit dominance. Several alternative explanations help to account for the makeup of the charter school market. However, because these alternative explanations have limited capacity to explain the charter market, contract failure appears to be the most significant factor driving nonprofits’ relative success.

The most obvious alternative explanation is state law: Most states prohibit for-profit corporations from receiving charters. This explanation is less promising than it initially appears because it cannot account for the infrequency of hybrid schools in most states. Only six of the forty-one states with charter laws statutorily prohibit hybrid schools. If legal restrictions on direct charter-holding were the only obstacle to greater for-profit presence in the market, then hybrid schools that avert legal restrictions but functionally approximate for-profit schools would be more common than they are.

Another possible explanation is that charter schools are simply unprofitable. Reliable data on the profitability of charter schools is extremely difficult to find, so it is hard to assess the premise of this explanation directly. But such data would illuminate little. It goes almost without saying that it’s hard to turn a profit in for-profit charter schools. If it weren’t, there would be more such schools. The real question, then, is not whether for-profit charters are unprofitable, it’s why. The contract failure theory I have espoused allows us to move beyond the somewhat tautological observation that for-profits are not particularly profitable and begin to understand the reasons why this is so.

A related explanation is that charter schools are unprofitable because their revenues are fixed. The amount of money that follows each child to a charter school remains static, no matter how much the school spends on its operations. For-profits cannot make money, the argument goes, because they cannot charge enough. This argument has the same problem that the general unprofitability argument does: It fails to explain why nonprofits are capable of operating with fixed revenues while for-profits are not. More important, pricing is a function of both quality and revenue. Even if revenues are fixed, for-profit schools can simulate price increases by lowering quality. As long as revenues exceed costs, profits are possible, regardless of whether revenues are fixed. Imagine, for example, a for-profit school situated in a neighborhood where the conventional public schools are so bad that the for-profit school can attract a large student body merely by
putting a roof on the building. Such a school could easily turn a profit by running a barebones operation just slightly better than the local conventional public school.

It is possible that, given fixed revenues, the return to capital on even bare-bones for-profit charter schools is simply too low for investors. One of the major differences between for-profits and nonprofits is that for-profits need to compete not only with other charter schools but also with other investment opportunities, and for-profit schools may be unable to match these opportunities with fixed revenues. Fixed revenues need not supplant contract failure as the major explanation of nonprofit dominance, however. Fixed revenues are problematic only if costs are high. And contract failure theory helps to explain high costs with reference to the challenges of overcoming the skepticism of parents, regulators, and donors.

III. EXPLAINING FOR-PROFIT EXISTENCE

Given that monitoring costs in for-profit charters are so high, it seems strange that any parents, agencies, or donors would prefer for-profits. Since nonprofits control agency costs more efficiently than for-profits do, why do for-profits exist at all? The answer lies in for-profit schools’ superior ability to raise capital and to exploit economies of scale.

There is ample evidence to suggest that most for-profit charter schools operate at a larger scale than nonprofit schools do. Scale in schools can be measured both in individual school size and in school network size. The available evidence indicates that for-profit charter schools are larger than nonprofits along both dimensions. Heath Brown recently measured and compared for-profit and nonprofit charter school scale on both dimensions in a four-state survey. Brown found that for-profit charters tend to operate larger individual schools and to centralize administrative functions into network offices at a higher rate than nonprofit charters do. Other evidence supports Brown’s finding on network size. The ASU study surveyed 436 for-profit charter schools across the country in 2004-2006 and found that 336—about seventy-seven percent—were operated by firms that had networks of ten or more schools. The data in the ASU report must be taken with a grain of salt, since they probably omit a disproportionate number of small for-profit operators. But data from a more detailed report on Michigan charter schools also support this story. In 1999-2000 in Michigan, 24 of the 122 for-profit schools were operated by firms that controlled one or two schools, 36 were operated by firms with three to six schools and 62 were operated by firms with more than six schools. Three firms together operated a total of 55 schools. The ASU report also supports Brown’s findings on individual school size. The ASU authors found a strong tendency among management companies with large networks to operate individual schools larger than the national charter school average.

For-profit schools’ ability to run large networks stems largely from their superior access to capital, the most essential ingredient in economies of scale. For-profits do not have to rely on the kindness of individual donors, but can turn instead to the more tried and true methods of the profit motive and securities markets. It is difficult to say exactly how for-profit schools go about finding investors, or who these investors are, because the great
majority of for-profit charter management companies are privately held, and therefore do not file periodic reports with the SEC.

The experience of Edison Schools is perhaps illustrative, if unusually dramatic. During its early stages, Edison Schools raised hundreds of millions of dollars from investors in private transactions (the precise amount is unclear), and notified the SEC that it would attempt to raise about $150 million in its initial public offering in 1999. The company eventually reached a total market capitalization of more than $700 million. A private equity firm purchased the company in late 2003 for around $100 million. There do not appear to be any nonprofit charters that have raised anywhere near as much money as Edison.

Using capital to run large-scale operations makes some for-profit charters competitive with conventional public schools and nonprofit charters. In these cases, the inefficiencies of monitoring and the efficiencies of scale interact to produce an equilibrium between for-profit presence and nonprofit dominance at which the advantages of scale make some parents, governments, and donors indifferent between nonprofit schools offering low monitoring costs and for-profit schools offering high monitoring costs along with the benefits of scale. Since the equilibrium point we observe currently involves only a fourteen percent market share for for-profit schools, we may deduce that the efficiencies of scale in for-profits are small relative to the inefficiencies of high monitoring costs.

We can make this abstract discussion more concrete with a simple example. Imagine a parent whose son wants to play high school football. The parent is contemplating sending her son to a small nonprofit charter high school. If only a handful of boys are interested in playing football, or if the school is unable to afford a football field because it cannot spread the cost sufficiently across its small student body, then the school won’t be able to field a team. If the parent’s choice is between this nonprofit charter school and a for-profit charter school of similar size, she’ll send her son to the nonprofit school, because of the risk that the for-profit school will imperceptibly cut quality to benefit shareholders. If, however, the for-profit school enrolls enough students to field a football team and is able to build a field, the parent might rationally run the risk of sending her son to the for-profit school to take advantage of the football team.

As this example demonstrates, there are many potential economies of scale in education, both on the individual school and network levels. Large individual schools experience economies of scale because they can spread fixed costs—such as football fields, administrators, libraries, classrooms, driver’s education practice ranges, and cafeteria equipment—across many students. They accomplish this by using these resources intensely. A school with only 30 students, for example, may require access to 5000 library books to cover the full range of subjects necessary for a complete education but will use these books very lightly. A school with 500 students may require the same number of books to cover the same topics, but may use these books much more frequently without seriously diminishing their value to each student. Large schools may also allow more subject specialization among teachers.
Large school networks experience similar economies of scale. For example, they allow managers to centralize decisionmaking, eliminating the need for each school’s principal to invest time in becoming fully informed and weighing the options for every decision facing the school. Networks can also centralize data collection, reporting, and accounting, which may become major burdens for small schools. Brown’s study observes strong tendencies in for-profit charter networks to centralize these kinds of functions. Additionally, networked schools can centralize purchasing, perhaps obtaining volume discounts from suppliers. Finally, large networks diversify risk and provide additional security for creditors. If one school in the network has a bad year financially, another may have a good year, and one school may be used to secure loans to open another.

There is undoubtedly a tension between the efficiencies gained from scale and the harmful effects of large schools on achievement. Large classrooms or perhaps even large schools arguably hurt achievement. Indeed, one of the more subtle goals of charter policy is to reduce school size. How for-profit schools balance this concern against the advantages of scale is not clear. It appears, however, that the benefits of scale are sufficiently large to create spots for at least a significant minority of for-profit schools.

IV. ORGANIZATIONAL FORM AND THE GOALS OF CHARTER SCHOOL POLICY

Besides telling us how charter school operators choose between for-profit and nonprofit forms, agency costs theory can also tell us how charter schools’ choice of form affects the schools’ abilities to achieve some of the more complicated goals of charter school policy. Charter advocates have cited a wide variety of possible advantages and efficiencies in support of charter schools, including the attraction of additional resources; circumvention of costly regulations and commitments such as union contracts; desegregation; improved opportunities for parents and teachers to participate in school governance; and a variety of benefits that stem from output-based accountability (such as efficient input use, satisfaction of parents’ preferences for idiosyncratic curricula, and greater experimentation and innovation).

In this Part, I will consider the way that charter schools’ choices of organizational form impacts three of these policy goals: resource attraction, parent and teacher participation in governance, and the benefits that stem from output-based accountability. I focus only on these three goals because there is no reason to believe that for-profit and nonprofit charter schools differ meaningfully in their abilities to achieve the other goals I have described. This is true, specifically, with regard to desegregation and circumvention of costly regulations and commitments. Both for-profit and nonprofit forms have advantages, and whether a state should favor for-profits or nonprofits may depend at least partly on the state’s needs. In most states, nonprofits probably better serve the complex ends of charter school policy.

A. Resource Attraction

One goal of charter policy is to attract resources that otherwise may never enter the public education system. Nonprofits and for-profits differ significantly in the types of additional resources they attract. For-profits tend to attract private capital and investment.
Unlike nonprofit schools and conventional public schools, for-profits are capable of going into private securities markets to offer equity shares in schools. For-profit schools may also be capable of obtaining innovative forms of debt financing that nonprofit and conventional public schools cannot obtain. Nonprofit schools, in contrast, attract donations. As noted above, these donations may come in the forms of volunteer labor, undercompensated employee labor, or cash.

On balance, policymakers will find the resources attracted by nonprofit schools most helpful. There may be instances in which private for-profit firms’ capital-raising abilities may be useful. If for some reason a district is short of capital—enrollment is growing rapidly, for example, or the district cannot get voter approval for its bond initiative—the possibility of authorizing hybrid charter schools to build new schools with private financing may be attractive. But these instances are rare. Many conventional public school districts have good credit and can obtain funds through low-cost or even tax-preferred bonds. Authorizing for-profit charter schools to issue securities at rates that reflect inferior tax status and the risks associated with an immature industry seems a rather roundabout way of dealing with capital shortages. On balance, therefore, nonprofit schools probably attract more useful resources to public education than for-profit and hybrid schools do.

B. Localized Governance

Many charter school advocates have argued for charter schools as an antidote to large-scale operations and as a way of empowering local teachers and parents to become involved in school governance. In spite of their efficiency in reducing costs, large schools and large networks of schools have been said to cause a variety of ills. Large schools may alienate and disorient students, and large, centralized networks may remove authority for crucial decisions like curriculum design from those best positioned to make them. Centralized governance may also discourage parents from participating and volunteering in their schools. Some charter advocates have argued that charter schools may be a solution to this problem if they can reduce school size and encourage parents and teachers to reengage in the governance of their schools. For-profit charters clearly are not as effective as nonprofit charters at achieving this goal. As noted above, for-profit schools’ most viable business model is to operate larger networks and larger individual schools than nonprofits do and to centralize decisionmaking. Even for-profit charters, however, tend to run smaller individual schools than conventional public schools.

C. Output-Based Accountability

Many charter school theorists have pointed to a variety of benefits that may derive from charter schools’ accountability for outputs rather than inputs. The value created by output-based accountability may take a variety of forms. For example, market pressure may push charter schools to satisfy idiosyncratic parent and student preferences for particular novel curricula. The National Study of Charter Schools found that “the primary reason for founding charters was to realize an alternative vision for schooling.” These idiosyncratic curricula improve efficiency by allowing parents and students who prefer these curricula over more traditional methods to derive more satisfaction from the same
level of government spending. Even if distinctive curricula do not improve students’ performance on standardized tests, they make parents and students happier and more satisfied with schools.

Another possible benefit is the incentive introduced by market competition to use inputs efficiently. Milton Friedman provided the first forceful articulation of these benefits in 1955, and numerous other proponents of school choice have since made similar arguments. John Chubb and Terry Moe argue that choice forces inefficient schools to close and allows efficient schools to thrive through “natural selection.” Before the No Child Left Behind Act (and perhaps even now) conventional public schools may have had inadequate incentives to hire capable teachers, for example, because the negative consequences may not have been significant (at least in a market-accountability sense). Ideally, market competition and the threat of charter school closure forces more efficient input use by charter schools.

Output-based accountability may also allow charter schools to innovate more efficiently than conventional public schools. Besides producing value by catering to idiosyncratic preferences, charter schools may produce value by creating gains on widely agreed upon standard measures of success because charters have both the regulatory freedom and the incentive to innovate. Though some early charter school theorists argued that regulatory freedom alone could produce innovation and experimentation, many theorists since have rooted these arguments in the pressures that output-based accountability produces. In Chubb and Moe’s account of the benefits of charter schools, for example, regulatory freedom improves charter schools’ ability to innovate and discover more efficient methods primarily because regulatory freedom makes schools more responsive to market pressures. They argue that only schools “unconstrained by bureaucracy” can change quickly enough to meet parents’ and students’ preferences.

All of this is worth thinking about because there is reason to believe that for-profit schools are slightly more responsive than nonprofit schools to the pressures of output-based accountability. Many early nonprofit theorists argued that in the absence of a profit motive and monitoring by interested shareholders, nonprofits would be less responsive to market pressures. If these theoretical predictions are true, policymakers may want to encourage for-profit schools, rather than nonprofit schools, to take advantage of the myriad benefits of output-based accountability.

These theoretical predictions favoring for-profits’ responsiveness to output-based pressures have not been strongly supported by empirical research, however. In the last twenty years, the relative efficiency of for-profit and nonprofit firms in responding to market pressures has been the subject of an extensive empirical debate, much of it focusing on the healthcare industry. Although some research has suggested that nonprofit firms are less likely than for-profit firms to structure managers’ pay in ways that directly incentivize efficient behavior, and other studies have found a positive correlation between for-profit organizational form and efficient input use, this body of research is ultimately inconclusive.

356
Draft — do not distribute!
Charter school theorists, as well as many state legislatures and charter regulators, have failed to recognize the important role that organizational form plays in charter schools. Although most states prohibit for-profit organizations from directly holding charters, only a handful of states acknowledge that the means for circumventing this restriction—hybrid contracting arrangements—are simple and widely used. Charter school regulators are also seemingly unaware of the importance the nonprofit role has played in ensuring charter schools’ success. Most charter school regulation emphasizes direct bureaucratic and parental monitoring, even though the nonprofit form may have done as much or more than these forms of monitoring to ensure charter schools’ efficiency. Additionally, charter school research has inadequately explored the consequences of organizational form for some of the more complex goals of charter school policy. Though regulators and researchers may have been aware that for-profit schools presented some risk for abuse, there has been no serious attempt to assess the impact of organizational form on goals such as resource attraction, localized governance, and responsiveness to output-based competition.

A number of policy reforms would make charter school regulation more responsive to the problems and opportunities posed by for-profit and nonprofit forms. First, while almost all state legislatures have already prohibited for-profit schools from directly holding charters, those states that have not should do so. Furthermore, legislators and agencies should close the loopholes that allow for-profit entities to operate charter schools by signing management contracts with nonprofit charter-holding entities. Most state legislatures, by prohibiting for-profit entities from holding charters directly, have made a judgment that for-profit charter schools are bad policy. My analysis suggests that this judgment is sound except in a limited set of circumstances. State legislatures should take their own judgment seriously by prohibiting for-profit entities from functionally achieving charter ownership through artful paper shuffling.

There is no reason to distinguish between direct for-profit ownership and hybrid arrangements that functionally emulate for-profit ownership. The nonprofit entities that contract with for-profit management firms in hybrid schools are unlikely to meaningfully check the for-profit firms’ tendency to exploit high monitoring costs. The nonprofit boards in hybrid schools are often controlled by the for-profit management firms. And even if the nonprofit boards begin by bargaining at arm’s length, they inevitably come to occupy extremely weak bargaining positions after the for-profit management firms completely take over their schools. In such instances, ending the contractual relationship means shutting down the school entirely.

Even if the majority of state legislatures are wrong and for-profit schools are actually good public policy, hybrid arrangements are still suboptimal. If for-profit schools are desirable, why force them to jump through the hoop of hybrid management structures? If state legislators seriously believe that for-profit schools make sense, the solution is not hybrid structures, it is direct for-profit ownership.
A prohibition on hybrid and nominally for-profit charter schools should perhaps be tempered by a couple of exceptions. There might be an exception for instances in which the for-profit firms are being used to remedy capital shortages in public schools. Such an exception would be difficult to craft. It might require, for example, a for-profit school’s application to be sponsored by a conventional public school district, or a public school district to prove to the chartering agency that the district is facing some capital crisis for which empowering a for-profit school is the best solution. The prohibition on hybrid and nominally for-profit charter schools could also create an exception for hybrid arrangements that grant for-profit companies only a small role in the management of schools. Such an exception might prohibit any contract between a nonprofit charter-holding entity and a for-profit firm that grants more than a certain percentage of the school’s annual revenue to the for-profit firm for management services.

Second, if state legislatures refuse for whatever reason to prohibit nominal and functional for-profit control of charter schools, then state legislatures or administrative agencies should place indirect limits on management firms’ ability to cut costs and increase profits. Charter regulators already monitor charter school finances relatively vigorously. These regulators could also use their administrative lawmaking authority to fashion and enforce rules limiting the profits that even for-profit management firms can make. Alternatively, agencies might mandate certain inputs—such as teacher or curriculum certification—or limit the amount of money that for-profit schools may spend for purposes other than the procurement of inputs to be used on a school’s premises or at centralized administrative offices. These kinds of regulations cannot reproduce the culture of nonprofit organizations that Hansmann and others have identified as being important to nonprofits’ effectiveness. But they may curb the most egregious attempts to cut quality in ways that parents and governments cannot observe directly.

Third, legislators and regulators should, to the extent possible, close the loopholes in existing monitoring regimes. Some of the failures of charter school monitoring are inevitable. Tests, for example, are always limited and they will always imperfectly match the highly complex goals of education. But many of the monitoring failures that have produced nonprofit dominance are the results of flaws in charter policy design and implementation, rather than inevitable obstacles that can never be overcome. For instance, policymakers should ensure that charter-monitoring agencies have adequate staff and resources to gather whatever information is available and disseminate it to the public. These agencies can consistently gather test data, financial data, and enrollment figures, and can make these data available to the public, thereby helping parents more effectively monitor charter schools themselves.

Finally, enforcement authority must be clearly delineated. Everyone—regulators, operators, and parents—should clearly understand which agency can force accountability and on what terms. Because closing a charter school often requires agencies to overcome significant public opposition, rules regarding achievement should be stringent, clear, and permit little opportunity for dispute.
CONCLUSION

My goal in this Note is not to defend charter policy against some set of criticisms or to argue that charter schools should be enabled, funded liberally, or otherwise encouraged. Whatever benefits charter schools may produce must be weighed against a substantial set of costs, none of which I explore here. My argument is simply that agency costs theory illuminates a number of interesting questions and problems in charter school regulation, some of which legislators would be wise to address.

This Note makes three main points. The first is descriptive: Nonprofits have come to dominate the charter school industry because they mitigate the agency costs that stem from parents’ and governments’ inability to monitor charter schools effectively and to hold them accountable. Nonprofit schools are able to perform this function because (1) legal restraints on profit distribution reduce the schools’ ability to exploit parents and governments’ inability to monitor, and (2) nonprofit charter schools tend to attract capable administrators and teachers personally devoted to charter schools’ missions. For-profit firms maintain a substantial presence, however, because they can raise capital more efficiently than nonprofit firms can. Sometimes the efficiencies of running large-scale operations with this capital outweigh the inefficiencies of high monitoring and agency costs.

The second point is also descriptive: Nonprofits may be slightly better than for-profit schools at achieving some of charter school policy’s more complex goals. In particular, nonprofits may attract more valuable resources and improve teacher and parent participation in governance.

The final point is prescriptive: Charter school regulation must more tightly control for-profits. For most of the charter school movement’s history, the nonprofit form has compensated to a large extent for the limits of the more direct forms of accountability that tend to consume policymakers’ and researchers’ attention. Given the substantial number of for-profit charter schools, regulators should limit hybrid for-profit-nonprofit arrangements, curb for-profits’ ability to cut quality in imperceptible ways, strengthen existing bureaucratic monitoring regimes, and more clearly allocate authority to enforce accountability. Organizational form has played an important role in shaping America’s charter schools. Policymakers would be wise to take it more seriously.
§ 6.3. Public Law and Accountability


I. INTRODUCTION

Professor Minow identifies accountability as a fundamental consideration when examining the allocation of various tasks to public and private actors. She writes that “[a]ccountability in this sense means being answerable to authority that can mandate desirable conduct and sanction conduct that breaches identified obligations.” Professor Minow concludes that greater private involvement in traditionally public sectors is both inevitable and in many respects desirable, but she also has significant concerns about the loss of public accountability. She acknowledges that competition and the private market can provide a form of accountability but observes:

Privatization of public services soared precisely when major corporations engaged in unfettered private self-dealing and one major religious group reeled from scandals, cover-ups, and mounting distrust among the faithful. This coincidence in timing should be all the reminder anyone needs of the vital role of public oversight and checks and balances.

This quotation, in our view, captures the fundamental problem with Professor Minow’s article. She moves directly from making observations about flaws in private markets to drawing conclusions about the importance of maintaining public sector influence in various settings. This line of reasoning is inadequate. In discussing the merits of public and private action, the analysis must be relative; flaws in private markets, significant though they may be, pale in comparison to the flaws associated with public provision or even public oversight of private actors. To conduct such a comparison, we first review how public and private provision of goods and services differ with respect to accountability. Using insights about the advantages of private provision from this discussion, we then examine the merits of public oversight of private actors. Public oversight, while appropriate in some cases, is not critical to a sensible privatization strategy.

This Commentary identifies two different aspects of accountability—accountability within an organization and accountability of the organization itself—and elaborates on the significance for accountability of the public or private nature of the provision of various goods and services. The first aspect of accountability is concerned with how the organization can motivate its agents to make decisions that further the organization’s objectives, rather than the agents’ self-interest. We discuss reasons why private firms are generally better able than their public counterparts to motivate their agents to pursue their organizations’ objectives. The second aspect involves the objectives that an organization should, from a social perspective, pursue. Private organizations are better able to pursue

---

the profit objective, and indeed most objectives, than are state agencies. Private actors face discipline that public actors do not. Moreover, there is a danger that public actors, even if potentially more successful at pursuing a particular end, will pursue socially undesirable ends because of political self-interest.

This Commentary proceeds as follows. In considering the issue of accountability within organizations, Part II examines the relationship between profit maximization and the organizational form, showing that private for-profit organizations are generally superior to governmental structures in their capacity to motivate agents to create profit. This, in turn, implies that when the profit motive creates socially optimal incentives, there is little need for direct government participation in the market. Part III turns to the question of why the profit motive might be undesirable from a social perspective in some settings, reviewing situations in which the market is inadequate along some dimension. When this is so, a direct public presence in the market may be desirable. But it may also be the case that government action would be no better in such situations, and would possibly be worse. Part IV relies on principal-agent theory to analyze the failures of public provision. The problems with private provision often imply severe deficiencies with public provision. Moreover, political pressure may pervert the goals that state actors pursue. Part V acknowledges that in some cases, public intervention is appropriate, but argues that the private sector and the profit motive should be relied upon when possible.

Part VI turns to the question of accountability. We consider two classes of reasons why accountability could be important in the provision of goods and services. There could be instrumental reasons: public or private agents that are held accountable to some constituency are more likely to behave in socially desirable ways. But because principal-agent problems are generally better addressed by private, profit-maximizing organizations, as suggested in Parts II to IV, accountability provides a weak rationale for direct public intervention in markets. The same merits of private markets also argue against significant public oversight of private providers. We examine the substantive merits of public oversight mechanisms, pointing out that these will be expensive and often either irrelevant or counterproductive.

Second, there could be non-instrumental values associated with accountability: it may be important in a democracy to have certain types of actors in certain sectors held accountable to the public regardless of welfare effects. It is doubtful, however, that this class of concerns necessarily justifies extensive public intervention. In many settings, self-interest often motivates private actors to behave as though they are public-spirited. Part VII sets out our conclusions.

II. WHY PRIVATE FIRMS DO BETTER THAN THE STATE IN PURSUING PROFIT

In this Part, we briefly review economists’ views of the theory of the firm and apply this theory to the question of public versus private provision, showing that important disciplinary factors are present in the private context that do not exist in public settings. Private firms are therefore typically more successful than the state in pursuing profit. In the subsequent Parts, we review what economics has to say about the social welfare
effects of the pursuit of profit, outlining both the socially desirable effects of the profit motive as well as cases in which the pursuit of profit alone is socially undesirable.

A. Pursuit of Profit in Private Firms

Private firms organize themselves to maximize profits. Professors Alchian and Demsetz stress the importance of the residual claim to profits in providing incentives for efficient production. A key feature of production within a firm is the joint nature of some work, like piano moving, which makes it difficult to reward workers on a contractual piecework basis. Instead, the firm operates through monitoring. Employees in a firm are supervised and rewarded formally and informally for good work. But the question arises: who monitors the monitors? Alchian and Demsetz rely on the allocation of the residual claim to the owner of the firm, or to its shareholders, to provide incentives for monitoring the firm’s production.

Professors Jensen and Meckling observe that the incentives provided by the residual claim are weaker when there is more than one shareholder. If there are several shareholders, the manager will not have the full residual claim, which will dampen her incentives to make profit-maximizing decisions. There are agency costs associated with the separation of ownership and control that are manifest in three ways: managers may take costly steps to commit themselves to avoid wasting corporate resources; shareholders may take costly steps to monitor managers; and some wasteful decisions occur that are too costly to eliminate.

But while agency costs persist, a variety of mechanisms exist that can reduce them. Executive compensation contracts, the risk of takeover, the value of managerial reputation, and the fear of failure if a firm in a competitive market wastes resources all serve to discipline managers. Many legal features of the for-profit corporation, such as the allocation of the residual claim to shareholders rather than, say, to customers, also encourage the maximization of profits and firm value.

One manifestation of these incentives to pursue profit is that a private firm has incentives to undertake tasks within the firm if doing so is efficient relative to contracting out. For example, Professor Coase suggests that the firm structure emerged because of the advantages of contracting within the firm, as opposed to arm’s-length contracting. Transactions within the firm are accomplished through simple commands, which entail lower transaction costs than entering into a consensual contract for every transaction; however, the disadvantage of entering into transactions within the firm is that the price mechanism is suppressed. Subsequent theorists have stressed the relative costs of different forms of opportunism under external contracting and internal production regimes. In the case of external contracting, there are incentives for the outside contractor to engage in pecuniary forms of opportunism by chiseling on features of the contract that are difficult to specify, monitor, or enforce. But when production is moved in-house, there are incentives to engage in nonpecuniary forms of opportunism, such as consumption or slacking on the job. Private firms trade these considerations off in arriving at the boundaries of the firm.
In contrast, the pursuit of profit will be less successful in governmental organizations. Governmental organizations have residual claimants, but this only tenuously affects incentives. The government is the residual claimant of its operations, and the relevant jurisdiction’s citizens can be thought of as the residual claimants on the government. The government can be the residual claimant either explicitly or implicitly. It is the implicit residual claimant wherever it carries on operations directly—that is, not through a state-owned enterprise (SOE). Where there is an SOE, however, the government is explicitly the residual claimant. In these enterprises, the government is generally the sole shareholder in a corporation. Such SOEs are common in many parts of the world other than the United States, although recent years have seen a worldwide trend toward privatization.

Allocating the residual claim to the government, however, does not create incentives in any given agent to maximize value. The question thus becomes, who has the residual claim in the government and therefore the incentive to monitor government agents? In a democracy, citizens do. They can realize this claim in a number of ways. Most commonly, the government uses proceeds from an enterprise to fund a different government program. But the government also might send checks directly to its taxpayers, as some North American governments have recently done, or reduce taxes.

Since there are residual claimants in government activity, why is it generally accepted that public and private provision of goods and services differ in their operation? One obvious and important reason is that the profit motive is socially inadequate in a number of contexts, and these contexts are often those in which governments intervene. But even where the profit motive does provide socially desirable incentives, the allocation of the residual claim to citizens through the government does not provide the same incentives to maximize profits as does the private sector allocation of the residual claim to shareholders, for several reasons.

One reason is that, unlike the individual residual claimant in Alchian’s and Demsetz’s model, there are potentially millions of claimants in the government’s operations. Citizens will thus remain rationally apathetic. The difference between this situation and that of many widely held public firms, however, may be easy to overstate. Neither a shareholder with the right to one-millionth of the residual profit of a corporation nor a citizen with an implicit ten-millionth share in a state enterprise is likely to take a keen interest in monitoring. The number of residual claimants in public enterprise is a factor contributing to the superiority of private enterprise in pursuing profit, but it is not an overwhelming one.

Another reason why allocating the residual claim to state enterprise to citizens is generally not as effective as allocating the residual claim to shareholders is that no market for corporate control in public enterprise exists. In the corporate context, raider corporations will monitor a firm’s performance and can make unsolicited takeover bids if the firm is underperforming and the raider thinks it can do better. Alternatively, dissident shareholders can seek to gain control in a proxy battle. The threat of such a takeover or
proxy battle disciplines managers in a widely held corporation, but no such threat exists for SOEs or government enterprise generally, short of a general change of government, which deficiencies in the operations of a single agency are unlikely to precipitate.

Providing incentives through executive compensation is also more difficult in the public context. Market-based compensation, such as stock-based compensation—whether through stock grants, stock acquisition programs, or options—helps provide incentives for managers in the private setting to maximize profits and firm value. In the context of public provision of goods and services, by contrast, there is not the same opportunity to provide market-based incentives, since there typically is no market for the securities of a particular government enterprise. This deficiency reduces the ability of public entities to provide incentives directly through compensation contracts and also weakens the managerial labor market: the absence of information through capital market signals about managerial performance undercuts reputational motivations, since good work is less easily recognized by outsiders when there is no secondary market in an enterprise’s financial claims. While incentive compensation is possible in the public context, the absence of market-based compensation makes such compensation less effective.

Another significant difference between public and private enterprise is that competitive product markets may not discipline managers in the public sector as well as they do in the private setting. Managers of corporations depend on capital for their corporations to survive. Corporations obtain capital either through retained earnings or from the capital markets through new equity or debt investment. Competitive discipline requires a firm not to waste resources; if it does, it will eventually fail (that is, be wound up): retained earnings will disappear and new investment will not be forthcoming. In contrast, public enterprises often do not face a “hard” budget constraint. Rather, governments have access to capital through their taxation powers and may use these monies to fund operations, even if those operations would not survive in the private setting. This risk is present when governments supply goods and services directly, but also exists for SOEs. It is common for governments explicitly or implicitly to guarantee the debts of SOEs, which in turn allows SOEs access to capital markets that they would not otherwise have. Ontario Hydro, for example, racked up the largest debt of any corporation in Canadian history through guarantees of its debt by the Ontario government. The risk of liquidation may be remote for government enterprise, which in turn undermines the incentives provided by competition.

The risk of liquidation is also relevant in comparing the performance of private nonprofit firms and governments. Nonprofit firms as a definitional matter do not have residual claimants. Such corporations are legally barred from paying out dividends or otherwise rewarding investors. But nonprofit corporations remain exposed to the risk of being wound up. They must generate capital either from their operations or from donations. Poor performance by a nonprofit likely jeopardizes both sources of revenue, which in turn threatens the firm’s survival. Government enterprise, as noted, is less vulnerable to failure. The prospect of failure motivates agents in a nonprofit setting in a manner that suggests that these organizations also are less vulnerable to agency costs than is government enterprise.

364
Draft — do not distribute!
Empirical evidence strongly supports these theoretical suppositions about the ability of profit-maximizing firms to motivate agents better than governments can. In a recent comprehensive review of the empirical evidence on the privatization of SOEs, Professors Megginson and Netter report that research studies have almost always confirmed that private actors economically outperform public agents in the provision of goods and services. Productivity and profitability increase when SOEs are privatized. In developing countries, where SOEs have often been ubiquitous, few of their proclaimed social or economic objectives have been realized, and they have been an enormous drain on public treasuries. The termination of these enterprises would free public resources to fund dramatic increases in health care and education funding. Moreover, studies also reveal significant economic gains when governments contract out for the provision of goods and services. A review of various studies found that government-provided services, including airlines, residential waste collection, electric utilities, and postal service, were between thirty and ninety percent more costly than equivalent services provided by the private sector.

In summary, when the motive to maximize profits is socially desirable, the role for government should be minimal. Private for-profit corporations are generally better able to provide goods and services at the lowest cost and prices. In addition, incentives to innovate in terms of product quality are typically much stronger in the private sector. Moreover, private nonprofit corporations are also likely to outperform government because of incentives provided by the risk of failure. This is not to say that private corporations never suffer from agency costs. As Professors Jensen and Meckling point out, there will sometimes be residual agency costs that are too costly to eliminate. Sometimes there will be a failure of governance at private corporations, as recent events cited by Professor Minow involving Enron, WorldCom, and Tyco illustrate. Nevertheless, private provision is on average superior to government provision in pursuing profit. However, in many settings in which governments operate directly, the profit motive is inadequate. We examine reasons for this in the next Part.

III. SOCIAL WELFARE AND PROFIT MAXIMIZATION

A fundamental theorem of welfare economics suggests that, under certain restricted conditions, the outcome of competitive markets is Pareto optimal, meaning that no individual can be made better off without making someone else worse off. The pursuit of self-interest by competitive, profit-maximizing firms drives prices down to marginal cost, which in turn maximizes society’s gains from trade. To draw a general distinction between private values and public values is inappropriate; private, self-interested actions often serve the public interest well. While Professor Minow acknowledges that competition and markets can serve the public good, she states:

When governments work closely with companies organized to make profits, they open avenues for investment-based financing, such as venture capital, that demands an ownership interest. Such undertakings also risk serious conflicts between public and private interests. Voters may object if public dollars more visibly enhance the earnings of private investors than they palpably improve the quality of schools, prisons, or housing for the poor.
The exercise of monopoly power is a genuine concern, as we discuss shortly, but it is not at all clear why public and private interests diverge simply because private investors are seen to make profits. Food retailers make money, yet we entrust them with a socially indispensable role without apparent objection. Indeed, any profit realized by the private provider resulting from cost savings relative to public provision is a social gain. In many, perhaps most, cases, incentives that motivate private organizations to maximize profits also lead to significant public benefits, even though the motivation is self-interested.

It is true, however, that in several contexts the profit-maximizing motive leads to social costs. In these settings, it is possible that direct government intervention could improve on markets. Nevertheless, several sectors in which governments around the world participate exhibit no obvious market failures that require public provision. One need only examine the inconsistency of public intervention to draw such an inference. Why do governments run railways but not intercity bus lines or taxis? Sell electricity but not oil or gas? Sell liquor but not beer? Professor Minow focuses on particularly problematic sectors in which the optimal mix of private and public provision is unclear, such as education and prison services. But it is worth remembering that relatively few sectors engage the public values that Professor Minow identifies in the education and prison contexts.

In this Part, we review reasons why economists view the pursuit of profit as leading to social costs in some contexts, and how government intervention has arisen perhaps in response to these costs.

A. Externalities

A crucial reason why profit-maximization may lead to socially perverse incentives is the presence of externalities. Presumptively, a consensual trade between two parties increases the parties’ private welfare, since the parties themselves, if made worse off, simply would not trade. Trade may not increase social welfare, however, if the transaction adversely affects third parties. For example, a person living in a thin-walled apartment may hire a trumpeter to wake her at four a.m.—this increases the trumpeter’s and the customer’s welfare, but may decrease social welfare overall, owing to the harms to neighbors. Absent some constraints, the pursuit of profit may lower welfare.

Governments often participate in markets in which there are externalities. One example is residential garbage collection. If each resident on an urban street contracted individually for garbage removal, social welfare might be lower than if garbage collection were coordinated centrally: each resident might fail to account for the cost to her neighbors of a large truck rumbling down the street. A single collector minimizes social costs by minimizing truck traffic. Therefore, government intervention, perhaps by centralizing garbage collection through contracts with private providers, may be required to reach an optimal outcome.

Governments may also react to externalities by requiring licenses for the activities that create them. For example, a bad driver endangers herself when driving and may take this into account, but may not account sufficiently for the potential injuries she may cause to
others. Likewise, a fisherman may not account for other fishermen in deciding whether to release an undersized fish. Consequently, there are driving and fishing licensing bodies. However, a purely profit-driven licensing body may provide inadequate screening and enforcement because of the same externalities that may cause individuals to behave socially harmfully in a world without licensing. An analogous example is liquor provision. Governments may participate in the provision of liquor because of externalities caused by underage or excessive drinking. And there may be a concern that the profit motive would lead to the serving of liquor to minors or inebriates.

A more subtle example involves education. As Professor Minow discusses, education may have a dual purpose. On the one hand, it may serve to teach individuals vocational skills, the benefits of which are largely realized by the student. On the other hand, it may serve to inculcate a wide variety of values, which we will call “citizenship values,” in students. The benefits of citizenship may entail significant positive externalities for society as a whole. Conversely, society suffers if, for example, students are taught to value sedition. Leaving the choice of education to individuals thus runs the risk of market providers emphasizing vocational skills over citizenship. To the extent that learning citizenship values does not result in private benefits and thus would not be rewarded by the market, the pursuit of profit would diminish the citizenship aspect of education.

Another controversial area of government and private involvement that raises similar issues is prison management. If the purpose of managing a prison is twofold, to incarcerate prisoners and to rehabilitate them, there is a danger that profit-maximizing private prison providers will not pursue the latter: relative to the cost of simply incarcerating inmates, it is expensive to rehabilitate them. Additionally, there may be a perverse incentive to encourage recidivism, which would create more business in the future for the provider.

B. Public Goods

Closely related to the problem of externalities is the issue of public goods. These are goods the enjoyment of which is nonrivalrous and nonexcludable. The classic example is armed forces. If a country has an army, all citizens are protected from invasion. Private markets fail in the delivery of public goods because of free-rider problems: to the disadvantage of citizens paying for the public good, free riders will try to enjoy the benefits of the good without having to pay for it. Thus, public goods are underprovided by profit-maximizers because they cannot collect from their customers.

The armed forces provide one example of government provision of public goods. Police forces and the court system also have a public goods rationale, although this does not preclude private provision of security and adjudicative services. The government also treats a number of goods that are not strictly defined as public goods as functionally equivalent to public goods. It would be theoretically possible, for instance, to charge each driver in a residential neighborhood for every mile driven on local streets. The transaction costs of such an approach, however, overwhelm the benefits and, as a result, the state builds and maintains local roads. Fire departments blur the line between being justified on public goods grounds and being justified on externalities grounds. Again, it would be
theoretically possible to exclude certain residences from firefighting services, and thus provide these services privately. However, all residents of an area benefit from their neighbors’ having fire protection, and there is thus a danger of suboptimal provision of such services if left to the market. To avoid this result, the state often intervenes in the provision of firefighting services, although again, obviously not to the exclusion of private fire prevention precautions.

C. Imperfect Information

A different reason why markets may fail to maximize social welfare involves imperfect information. When consumers have difficulty maximizing their own welfare in making consumption decisions because they lack meaningful information, private markets will not work well. The following classification captures different information problems relating to three different kinds of goods: search goods, experience goods, and credence goods. Search goods are those whose quality the customer can evaluate prior to purchase. In the case of experience goods, the customer can only determine the quality of the good by using or consuming it. Credence goods are not easily evaluated either prior to or subsequent to their use. Markets can work well for search goods and experience goods, particularly through legal devices such as warranties and market features like reputations for quality, but may not work well in the case of credence goods. Profit-maximization would encourage private providers of credence goods to shirk on quality, since by definition the market will neither reward nor punish on that basis. Therefore, governments may provide the goods or services directly, or may regulate their provision. For example, pharmaceutical products are often credence goods. Governments provide certification services for these goods, not allowing drugs on the market without first vetting them for safety and efficacy or requiring medical prescriptions.

D. Paternalism

Related to the problem of imperfect information, governments occasionally intervene for paternalistic reasons. Minors, for example, are unable to enter into legally binding contracts out of a concern that they may not know what is in their best interests. This policy undermines the efficacy of some markets, particularly the market for education. It is difficult to borrow against human capital in many situations, because there is no collateral to collect and high-interest loans may be unavailable. In addition, there is no legal way for minors to go to capital markets to borrow for an education. This means that minors for whom education is a good investment (that is, virtually all minors), but who are unable to access informal sources of capital, like family, would be excluded from obtaining an education absent government intervention.

E. Monopoly

Monopoly is another source of market failure. A natural monopoly realizes economies of scale such that with each additional unit produced, the average per-unit cost declines. In such a context, competitive markets will not emerge; rather, the largest firm will eventually capture the entire market because of its cost advantages. There may be competition for the market, but there will not be competition within the market. A profit-
maximizing natural monopolist will make socially suboptimal output and pricing decisions: It will restrict output and increase price in order to maximize profits. In this manner, some customers who would be willing to pay the marginal cost of production will be priced out of the market—the monopolist forgoes some potentially profitable transactions to maximize its share of the gains from trade that does take place. In the face of such a natural monopoly and the social costs it entails, the government may intervene. Intervention may involve government provision of the monopolized good. For example, government agencies often provide utilities—telephone, electricity, and water service—because there are significant scale economies in these industries. Government intervention may be less stark, however, and may take the form of regulation of a private natural monopoly.

F. Distributive Concerns

There is another reason for government intervention that arises not because of market failures, but rather out of distributive concerns. Competitive markets create wealth, but also create income and wealth inequalities. To the extent that society has a preference for egalitarianism in the distribution of wealth, governments may provide certain goods and services. For example, the United States government has long provided food to needy families, initially by providing the food directly and later through food stamps. If redistribution is socially desirable, and if there are inefficiencies in relying exclusively on the tax system and cash transfers to redistribute wealth, government intervention in the provision of goods and services could be desirable.

IV. DOES THE COST OF MARKET FAILURES EXCEED THE COST OF GOVERNMENT FAILURES?

Private organizations are generally more effective than governments in motivating agents to act in ways that maximize a firm’s profits and value. This justifies the significant trend toward privatization of SOEs in many parts of the world outside the United States (which has almost no such enterprises). In some contexts, maximizing profits does not maximize social welfare, and there may consequently be a role for government provision in these settings. But even if there are imperfections in private markets, there are potentially more significant drawbacks associated with public provision. In our view, Professor Minow pays insufficient heed to the potential for government failure. We discuss two general problems. First, governments generally are less able to motivate their agents to pursue a given objective than are private parties. Second, governments are vulnerable to political rent-seeking, which perverts the ends they seek to pursue. We develop these points in turn.

A. Problems in Achieving Government Objectives

Government enterprises often fail to motivate their agents to accomplish desired objectives to the same extent as private parties. If the goal is profit-maximization, we outlined above a number of reasons why private, for-profit firms perform better than government enterprises. But even if the appropriate social goal is not profit-maximization, private firms have an important advantage over governments: private
corporations may not survive in the face of poor performance, while government enterprises generally do not face such a risk. Agents at nonprofit corporations have incentives to perform well to ensure their organization’s survival and retain their employment. Good performance results in higher operating revenues or more donations; bad performance could result in lost jobs because of the organization’s liquidation. Also, private corporations have stronger incentives to hire employees committed to the objective in question. Fearing failure, the nonprofit entrepreneur will seek to hire appropriate employees; governments without this threat have dampened incentives to do the same.

Aside from the generally superior capacity of private organizations to motivate their agents, it is important to recognize that the features that undermine the market often undermine public provision as well. Externalities are ubiquitous. And yet in many, indeed most, instances the state does not intervene. For example, if someone wears an ugly sweater, the public may suffer, yet nowhere are there regulations preventing ugly sweaters. (Indeed, the academy would be in jeopardy if there were.) The reason why the state more often than not declines to respond is that the cost of intervention exceeds the benefits. In the case of the ugly sweater, the externality persists in the absence of regulation because it is impractical to get all those in contact with the sweater in question to reveal the value of their displeasure and contribute to a fund to buy and destroy it. Similar difficulties render government intervention impractical as well. To conduct a correct cost-benefit analysis, the government would have to discern the value each citizen places on the destruction of the sweater. However, citizens generally would not have the correct incentives to reveal the truth: if their taxes did not depend on their answers, those wanting the sweater destroyed would overstate their displeasure with it; if their taxes did depend on their answers, they would have an incentive to understate their displeasure and free ride on the contributions of others.

More germane to our current discussion, sometimes the market failure that undermines private provision also hinders the ability of the state to motivate its agents to provide goods or services in an optimal manner. Consider education. Recall the concern that an individual may choose a course of education that is best for that individual but may not be good for society. If education were left to profit-maximizing firms, there would be a danger that skills training would be overprovided relative to citizenship training. Profit-maximizing educational providers may fail to motivate their agents to provide a socially optimal mix of skills and citizenship. Where the government provides education, it faces a similar challenge of establishing a set of organizational incentives that motivate agents to teach both skills and citizenship. Because of the combination of externalities and informational deficiencies, this is not an easy task. If teaching citizenship were verifiable, governments could simply require profit-making institutions to meet certain citizenship standards, perhaps through testing students, while competing as they saw fit over skills provision. The government’s role would be limited to standard-setting. However, if teaching citizenship is not verifiable, which seems likely, this option is unavailable. But an absence of verifiability that would undermine private provision of education would likely also affect public provision.
How would the government motivate its own agents to teach citizenship? Any incentive pay structure would skew incentives toward teaching skills, since by assumption such teaching can be measured, while citizenship cannot. Perhaps if teaching citizenship were observable, even if not verifiable, teachers could be monitored. But by whom? And what incentives would the monitors have? Alternatively, the government could provide education without relying on incentive pay at all. In this way, the bias toward skills over citizenship would diminish—teachers would obtain no obvious private benefit from deviating from the optimal mix. With such a compensation package, however, there may be little incentive to teach either skills or citizenship properly. This last arrangement is the norm in North America: public schools tend to provide teachers with very little in the way of merit pay. This approach may address the skills/citizenship bias, but may not provide much motivation to teachers.

Leaving aside externalities, consider only informational deficiencies in education. Consumers of education—particularly younger students—and their parents may be unable to evaluate the quality of the education they receive. If the quality were testable, standardized tests across schools might be appropriate—this would help consumers evaluate the quality of schools. However, there is no reason that private testing or ranking services could not emerge. Schools that fail to submit their students to testing, or to give information to institutions for ranking, would presumably suffer in the marketplace.

Suppose, however, that testing fails to capture important elements of education provision, such as interpersonal skills. Before the government should intervene, it must establish that it can actually do better than consumers in monitoring education providers. Even if government agents have comparative expertise in evaluating education, how are government agents motivated to monitor schools? Imprecise evaluations of schools by motivated students or parents (or at least a significant minority of informed parents) may be preferable to potentially accurate evaluations by relatively unmotivated government agents.

Professor Minow notes that perfectly competitive markets in the education sector require a number of features, like good information, that may or may not exist in practice. We disagree with her whether some of the factors she points to are truly necessary. For example, she suggests that “signaling [quality] requires networks of communication—and these are precisely what is jeopardized when informed and motivated parents leave public schools for private or charter schools.” But the very act of leaving serves as a signal. Networks of communication are not required to draw an inference about an empty restaurant’s quality. Even accepting that there are other infirmities in education markets, it is crucial to return to the question: as compared to what? Professor Minow neglects to scrutinize the potential failures of public education as thoroughly as she does those specific to private provision.

This last point generalizes. While the pursuit of profit fails to achieve optimal social welfare in some settings, government provision may also fail to achieve this goal, in part because agency costs often afflict governments more severely than private enterprise. As a general matter, this is because of insulation from the risk of failure, and in the particular contexts in which profit-maximization is not socially optimal, the reasons for market
failure often affect the government’s own capacity to motivate its agents. Assuming, then, that the government can be relied upon to choose the correct objectives where profit-maximization is inappropriate, it may not have the means to pursue these objectives cost-effectively. Infirmities in private provision may be preferable to those associated with public provision.

B. Distortion of Government Objectives

In addition to the problem of governments’ lacking the means to pursue socially optimal ends cost-effectively, there is also a significant risk that governments will not even attempt to pursue socially optimal objectives. Just as the profit motive or incentive may distort private sector priorities relative to broader notions of social welfare, political motives or incentives may distort both choice of policy objectives and choice of policy instruments in public sector decisionmaking. This is true of public provision or any of a host of other government interventions in private sector activity. Put another way, allocating to private parties the residual claim to profits can sometimes lead to social costs, but if the government assumes responsibility for the economic activity, the concern is that government actors will exploit access to the residual claim in a manner that could be even more costly.

A well-developed body of Public Choice literature seeks to demonstrate that the policy objectives of governments are often radically at variance with the normative prescriptions of welfare economics. The latter would limit state intervention to correcting for pronounced and reasonably well-defined categories of market failures such as monopoly, externalities, public goods, and information failures; by contrast, state intervention often in fact extends to any number of other objectives, including perverse redistributive objectives (for example, redistributing to well-endowed, concentrated interests). Even the more modest claim that, whatever the policy objectives of the government (efficiency, redistribution, communitarianism, and so forth), it is rational for the government to pursue these ends or objectives by the most efficient (that is, least costly) means, is also described as fundamentally misconceived. In Public Choice theory, existing policies that deviate from the prescriptions of welfare economics or other normative frameworks are rarely mere reflections of political ignorance or stupidity that will be corrected once economists or other social scientists point out to politicians the folly of their ways. Rather, prevailing policies represent a form of political equilibrium arrived at through collective decisionmaking processes whose classes of participants (politicians, bureaucrats, regulators, interest groups, the media, and voters) should not be viewed as involved in the common and selfless pursuit of some universally recognized set of public interest goals, but rather as players in a kind of implicit market involving intricate sets of exchanges among self-interested actors. That is to say, actors will be similarly self-interested whether they are acting in economic or political markets. Thus, for example, politicians seeking to attain or retain political office will find it rational to fashion policies that exploit various political asymmetries: between marginal voters (uncommitted voters in swing electorates) and infra-marginal voters; between well-informed and ill-informed voters; and between concentrated and diffuse interest groups facing differential political mobilization costs (collective action problems). Moreover, because of short electoral cycles, politicians will favor policies that yield immediate and
visible benefits and that defer costs or render them less visible (for example, by moving them off-budget). Bureaucrats will be motivated to promote policies that maximize their power, pay, and prestige. Regulators will seek a quiet life by accommodating the interests they are supposed to be regulating, and perhaps will also seek to enhance their prospects of employment in the regulated industry after their tenure as regulators (the “capture” theory of regulation). The media, to maximize readership or viewing audiences and thus enhance advertising revenues, will trivialize complex policy issues, sensationalize mishaps that may not reflect systemic policy failures, and turn over issues at a rapid rate with minimal investigative follow-up to cater to readers’ and viewers’ limited attention spans (rational ignorance).

One does not have to accept the most austere (or cynical) versions of Public Choice theory, or to dismiss entirely the relevance of ideas or values to public decisionmaking, to recognize that the theory has substantial explanatory power in many contexts. Contemporary examples of how Public Choice theory bears out in practice include farm subsidies and other forms of agricultural protection, steel tariffs, tax breaks for large corporations (for example, foreign sales corporations), and recurrent disaster relief in flood-prone areas that discourages mitigation strategies. Thus, both profit motives and political motives have the potential to distort policy objectives significantly, rendering any idealized presumption in favor of public provision utopian at best.

C. Market Responses to Market Failures

In contrast, competition in private settings may compel providers to adopt socially optimal objectives even when there are infirmities with the market. The medical services market provides an example of a market with information deficiencies. Many consumers of medical services are unlikely to be able to evaluate ex ante or even ex post whether the medical care they receive is adequate. Consequently, there is an incentive problem in that medical care providers may face temptations to shirk in different ways—through minimal effort, perhaps, or even by fraudulently misrepresenting themselves as trained professionals when they are not. It is not necessarily the case, however, that extensive government involvement is required to correct this market failure.

The U.S. health care system relies on private insurance and hospitals. Hospitals are often nonprofit institutions, which may suggest a perception that the unbridled profit motive could lead to inferior care. The temptation for profit-maximizing institutions to cut corners when providing health care to consumers unable to evaluate the quality of their care may induce customers to opt for nonprofit institutions, which have weaker incentives to shirk. Government intervention may not be necessary in this context, as the market is capable of evolving in a way that promotes the optimal private organizational form. Moreover, if private insurers are able to evaluate care, and if these insurers have some incentive to ensure that their patients get good care (perhaps because consumers or their agents, such as unions or employers, can evaluate an insurer’s performance over a large number of patients, even if they are unable to gauge the appropriateness of care in any given case), then the marketplace may provide adequate incentives for for-profit institutions.
D. Summary

We have argued not only that public institutions are particularly vulnerable to agency problems in many contexts in which markets fail in some dimension, but also that the danger of perverting the very objectives of provision is stronger when the state is the provider. It is true, of course, that governments may intervene in a sector for political reasons even when the state is not the key provider in question. However, the decision to allocate responsibility for provision to the government gives rise to a more acute danger of self-interested political involvement for two reasons. First, once the government has assumed control of an activity, the costs of political intervention in that activity fall. For example, politicians can be more discreet in setting goals for public actors than in interfering with private actors’ pursuit of profit.

Second, once an activity has been allocated to the state, an array of interest groups will develop that may be committed to a goal other than profit. Employees of the public provider will, for example, have an interest in lobbying for pursuit of employment security or higher wages, rather than profit. Before the government allocated a service to the public sector, there would be potential employees who would have some interest in an employee-oriented provider, but several factors may prevent them from pursuing this interest: they do not know for certain that they would benefit from such a policy, as they do not know whether they would actually be public employees; they do not have a pre-existing organization, like a public employees’ union, to lobby on their behalf; and they are sufficiently numerous that lobbying would be difficult to organize. The allocation of a task to the public sector therefore invites the participation of interest groups in the political process, which may not occur in the absence of such an allocation.

In short, where the pursuit of profit maximizes social welfare, governments should play a minimal role. Private organizations are better equipped to choose optimal means to realize this goal. But even when the pursuit of profit is socially costly, the appropriate role for the government may remain small. This is because private organizations, even if they are nonprofit, are better able to motivate agents, both in general and in those contexts in which there is market failure. Moreover, aside from having inadequate means of pursuing given ends, the government often fails to choose the correct ends. Rather than seeking to maximize social welfare, government actors may choose instead to maximize their own welfare. In other words, there is a “meta-agency” problem in that government agents may not seek to maximize the welfare of their principals, the public. This problem will have a greater impact after a particular activity has been allocated to the public sector, because such an allocation creates new interest groups.

V. DEFINING THE APPROPRIATE ROLE FOR GOVERNMENT

While the above discussion suggests a presumption against significant public provision, in some cases the costs of market inadequacies exceed those of government failures, and public participation in a market is therefore appropriate. But the discussion also has implications for choosing the optimal type of public intervention. Private, especially profit-maximizing, corporations are best able to discipline their agents. Government intervention should therefore rarely negate an important role for the private sector.
Rather, the profit motive should be constrained to the extent it causes social costs, and encouraged where it does not. A simple example demonstrates the point. Factories can cause pollution, but this does not invite nationalization of all factories. Governments instead may intervene in relatively targeted ways, such as by establishing pollution taxes or even more market-oriented, tradable pollution permits. We consider in this Part examples of appropriate roles for the government and the private sector in the context of different market failures.

Garbage collection involves externalities that result from the private choice of collectors: each resident fails to account for the cost to others of garbage trucks rumbling down the street. Centralized collection eliminates the cost of these externalities. This may require government intervention, but it does not require government provision of waste collection services: the government can simply contract out the task for given neighborhoods to private firms, perhaps through competitive tenders. Competition spurs firms to bid at low prices (setting aside collusive practices, which are not unknown in the sanitation business). There is competition for the market, not within the market, and the profit motive will spur agents within the successful bidding firm to provide the service efficiently. Unless the government can motivate its agents to behave more efficiently than private, profit-maximizing firms can—which, as discussed above, is unlikely—the truck traffic externality may warrant government intervention and centralization, but does not require the absence of profit motivation on the part of the firm (and its agents) that collects the garbage. It is not profit per se that creates poor incentives in the private garbage market, but individual choice, and where the profit motive can be preserved, it should be.

However, contracting out may invite problems of its own in some contexts. For example, Professors Hart, Shleifer, and Vishny, adopting an incomplete contracts perspective, view the objective function of a number of government activities as sufficiently complex or contingent that precise specification, by contract, legislation, regulation, or otherwise, of performance objectives and ready monitoring of the achievement of these objectives are infeasible. An extreme example is the formulation and implementation of a country’s foreign or defense policy, because complexity of objectives and unforeseeable contingencies render delegations of these functions to private actors highly problematic. In contexts such as this, it is unclear exactly what is being delegated. We have cited less extreme examples, such as inculcation of citizenship skills in education, earlier in this Commentary.

Hart, Shleifer, and Vishny conclude that the case for in-house government provision over external contracting is generally stronger when non-contractible cost reductions have large deleterious effects on quality, when quality innovations are unimportant, and when corruption in government procurement is a severe problem. “In contrast, the case for privatization is stronger when quality-reducing cost reductions can be controlled through contract or competition, when quality innovations are important, and when patronage and powerful unions are a severe problem inside the government.” Public ownership removes the tendency to engage in excessive cost reduction but replaces that tendency with a weaker incentive to engage in both sensible cost reduction and quality improvement. Which arrangement is superior therefore depends on which distortion is less damaging.
In a slightly later paper, Professor Shleifer identifies what he calls a rather narrow set of circumstances in which government ownership is likely to be superior. These are situations in which: “1) opportunities for cost reductions that lead to non-contractible deterioration of quality are significant; 2) innovation is relatively unimportant; 3) competition is weak and consumer choice is ineffective; and 4) reputational mechanisms are also weak.”

We do not necessarily disagree with any of these propositions. However, we reemphasize that private firms have access to a variety of mechanisms to motivate their agents to act in accordance with the objectives of the firm; government organizations and agencies, by contrast, lack access to such mechanisms. Internal agency problems are likely to be more severe when government organizations internalize functions, most particularly when problems of contractibility prevent clear specification of objectives or ready monitoring of adherence to these objectives. We have already argued, in analyzing whether circumstances ever warrant government intervention of any kind, that imperfections in the market in some settings, such as information problems in education, also afflict government provision, perhaps to the point where private markets are preferable. Further, in markets in which there are significant problems with contracting out, there are likely also to be significant problems with direct public provision. Such contexts are likely to drive government organizations or agencies to adopt elaborate command and control mechanisms for their internal agents that specify in detail inputs and processes, but not outputs or outcomes. Command and control mechanisms that seek to minimize costs in this way are likely to create minimal incentives for agents either to reduce costs or to engage in product quality innovations. Thus, in-house public provision may be as or more problematic in environments of poor contractibility.

The inability of minors to enter into legally binding contracts and thus to borrow money engenders a deficiency in the private education market. This may be another appropriate forum for government intervention, but, again, moving from inadequacies in market provision directly to public provision is inappropriate. The richness of intermediate solutions is something that Professor Minow rightly emphasizes. One obvious problem in this area is inadequate family financial resources. Rather than relying on a public school system operated by the government, it may be preferable to allocate funds to students to permit them to attend an educational institution of their choice. Educational vouchers are one option. Private schools in districts administering voucher programs will face competition in attracting students, which in turn will encourage the schools to motivate their agents. Without well-motivated agents, enrollment will decline, which could jeopardize the schools’ profits (or their existence, in the case of nonprofit schools). If the only deficiency in the educational market is the absence of borrowing capacity, there is a reason to provide vouchers to encourage efficient demand, while still allowing market mechanisms to determine the supply side of education.

Natural monopoly may also call for government intervention of a limited kind. The problem with natural monopoly is that profit-maximizing monopolists will sell at too high a price from a social perspective, and any move to introduce competition leads to social costs from inefficient scale. In response, the government could provide the good or service itself. In Ontario, for example, the government owns Hydro One, which owns and
maintains the transmission grid for electricity. But, for reasons of government inefficiencies discussed above, the government will face significant difficulties in motivating its agents to keep costs low. And if costs are not kept low, prices set merely to cover costs could conceivably rise above even a profit-maximizing monopolist’s price. A better way to deal with natural monopoly may be through private, profit-maximizing ownership of a regulated monopoly. An appropriate regulatory regime can limit prices while creating incentives for the firm to lower costs (performance-based regulation). This implies in turn that the profit-maximizing firm will seek to motivate its agents to lower costs, although every regulatory regime will have its own additional costs and imperfections.

Some markets give rise to significant distributional concerns. For example, some citizens have insufficient earning power to afford adequate housing. Society, for a variety of reasons, may choose to assist the poor in obtaining housing. But again, even if governments intervene, it is not clear that they should provide housing directly, even though they often do. Government agents will again have only tenuous incentives to build the best kind of accommodation in the most appropriate locations, whereas market suppliers can increase their profits by providing housing that beneficiaries of the public program want. It may be better for the government simply to transfer resources to those in need, perhaps through rent vouchers. The market would then reward those who provide optimal housing. Indeed, this move from direct provision to indirect provision through vouchers describes the evolution of food subsidies for the poor in the United States: the federal government initially operated warehouses of food for the needy, but later began allocating food stamps to be spent largely at the customer’s discretion in private retail stores.

In summary, profits tend best to motivate organizations and their agents. While profits sometimes lead to socially undesirable results, such market failures may also impede government intervention. Where intervention is appropriate, the government should be sensitive to allowing the profit motive to operate where possible.

VI. Accountability

Public values are at stake in a number of different sectors, but they do not always invite public provision. Private for-profit and nonprofit providers often vindicate these values while better motivating their agents than does the government. Professor Minow discusses a potentially important difference between public and private provision: public actors are held accountable in a variety of ways that private actors are not. She notes that “[p]rivatization creates possibilities of weakening or avoiding public norms that attach, in the legal sense, to ‘state action’ or conduct by government.” For example, public actors may have a duty to explain themselves, either through an administrative law duty to give reasons or through statutes granting citizens access to information. Public actors in the United States and Canada also face constitutional constraints that private actors do not. Professor Minow considers it fundamental that accountability be preserved in the presence of private participation in public sectors; this will in turn promote a variety of other values like democracy, equality, and pluralism.
It is true that there are fewer legal accountability mechanisms associated with private provision than with public provision. There are, however, strong reasons not to overstate the importance of the absence of certain legal accountability mechanisms in the private provision of goods and services. First, private actors are subject to other, often more potent, disciplines. Legal scholars tend to emphasize legal accountability mechanisms, but there are important economic sources of discipline on private actors, as we have discussed. Professor Minow acknowledges that “[p]rivate economic markets generate accountability through the operation of supply and demand, which tests the viability of ideas, products, and processes by their ability to attract and maintain a sufficient number of purchasers to meet costs and generate desirable profits.” But market-based accountability is not merely an incidental benefit of privatization; as previous Parts have explained, the benefits of privatization arise largely because of the accountability generated by private markets.

In addition, more than simply ensuring the survival only of the fittest, market forces often compel private firms to act as though governed by public accountability rules. To take a simple example, suppose education was provided by private firms. It is highly likely that private schools attempting to attract students would disclose significant amounts of information about their approaches and their success. One need only ride a Toronto subway to read of private vocational institutes advertising the success rate of their graduates in finding jobs. U.S. News and World Report annually ranks law schools in the United States with the cooperation of these schools. Private high schools also publicize their success rates in placing students in elite postsecondary educational institutions. Not only would the market require schools to provide a certain type of education to succeed, it would also require schools to disclose information about themselves. Private accountability mechanisms cause private actors to behave as though public-spirited.

Conversely, despite the importance Professor Minow places on them, it is not clear whether public accountability mechanisms work to discipline public actors. While Professor Minow worries that emerging arrangements “jeopardize public commitments to equality, due process, and democracy,” evoking by way of counterpoint civic republican conceptions of government, public accountability mechanisms are often light years removed from this ideal. Bureaucracies, for example, involve rigid job classification systems, lockstep and seniority-based systems of promotion, a lack of pecuniary or nonpecuniary reward systems for superior individual performance or for innovative or productivity-enhancing ideas, high levels of job security, rigid line-item budget constraints, budgeting cycles that entail forfeiture of unspent monies and reduced future budget allocations that create incentives to “move the money” at the end of the cycle—in general, a web of bureaucratic red tape driven by concerns over process and inputs and not outcomes. The agency problems that afflict public provision also afflict public accountability mechanisms.

Furthermore, public accountability exists for the most part through reliance on certain process values, like the right to hearings and the right to information. Substantive decisionmaking is generally only constrained to the extent that these procedures influence public debate, and only to the extent that public deliberation in turn influences the substance of government action. Because of this reliance on process, agencies are often
free to pursue a substantive agenda that may or may not advance the public interest. We note again significant Public Choice concerns over distortions in the choice of both policy objectives and instruments, including the disproportionate influence of concentrated interest groups, incumbency, campaign financing, and logrolling in the legislative process. The danger of interest group influence or regulatory capture often distorts the substantive direction of governmental policy, notwithstanding public accountability mechanisms.

Public oversight could in some instances largely undermine the benefits of privatization, which, as discussed, arise in part because of the insulation of actors from public influence. A recent example from Ontario is apt. The electricity market in Ontario has historically been characterized by extensive public participation and serious mismanagement. Recent reforms have included some degree of privatization and the introduction of competition in the electricity generation market. But in part because of uncertainty about continued public involvement, particularly because of Ontario’s continued ownership of the largest generator, Ontario Power Generator Inc., the private sector has been reluctant to invest in the market. This reluctance was prescient: in response to apparent public dismay over climbing electricity prices, the government recently set a ceiling on electricity prices, thus simultaneously undermining market incentives for consumers to conserve energy and incentives for potential producers to invest in generation capacity. The continued presence of significant public participation in the market implies the lingering specter of public oversight, which has seriously compromised the reform project.

This is not to say that public oversight is never appropriate from an instrumental perspective. The costs of market failures of various kinds will sometimes exceed the costs of public oversight. But keeping public participation to a minimum, as discussed in Part IV, is the best assurance against either simply misguided intervention by under-motivated government agents, or self-interested but socially undesirable intervention by ill-motivated government agents. The phrase “public accountability” can easily induce a slide to unreflective thinking that the more accountability and the more public it is, the better. In the real world of policymaking we should resist this temptation to indulge the nirvana fallacy.

To the extent that accountability is based on instrumental concerns, private accountability mechanisms compare favorably with public ones, as we have argued. But perhaps there are non-instrumental values associated with accountability. Perhaps it is simply important in a liberal democracy for a variety of due process rights to attach to the provision of certain goods and services. There are several responses to this concern. First, in several traditionally public contexts, it is very difficult to imagine what non-instrumental, democratic values are imperiled by private provision. The absence of public accountability mechanisms in household garbage collection, electricity generation, park maintenance, wine and liquor retailing, and road repair hardly seems to engage fundamental questions of democracy. Again, Professor Minow’s focus on problematic cases may divert attention from the many straightforward cases for privatization. But even in sectors such as education and child care services, in which democratic values, however defined, may be at stake, private provision advocates can argue that even if there
is a certain sacrifice of accountability, this loss does not trump the instrumental benefits brought about by privatization. Unfortunately, intellectual argument brings little guidance to this debate. Ultimately, we are left with a political tradeoff between the incommensurable value of greater welfare and whatever non-instrumental value is attached (by assumption) to accountability.

Second, even if accountability were to trump purely instrumental concerns, private provision leads to several accountability mechanisms. While these arise through the operation of self-interest and the market, the proponents of public provision bear the burden of explaining why private accountability mechanisms do not satisfy these non-instrumental concerns, given their superiority in other respects. Public provision advocates must identify the particular (non-instrumental) value that accountability vindicates and why private mechanisms are insufficient. Moreover, even if there is some non-instrumental reason why public actors ought to be subject to certain accountability mechanisms, this does not necessarily imply that private actors ought also to be subject to them. For example, constitutional restraints apply only to public actors in both Canada and the United States.

If private accountability mechanisms are deemed inadequate despite the preceding arguments, the state can simply impose legal accountability requirements on private actors. If, for example, markets are inadequate in compelling information disclosure, the government could compel it, as it does (rightly or wrongly) in the securities context. If due process is a concern, the government could impose these requirements on private actors. For example, landlords must fulfill significant notice requirements before evicting a tenant, and tenants often have rights to hearings in the same context. In addition, many market actors are subject to laws prohibiting discrimination in transacting business.

However, it is fundamental to acknowledge that the imposition of legal accountability or other constraints on the private sector may entail costs in terms of reduced competition, innovation, and flexibility, which may negate any advantages of private sector over public sector provision.

VII. CONCLUSION

Professor Minow emphasizes the importance of accountability. In this Commentary, we have shown that private actors are accountable in significant ways, even if accountability arises through the operation of the market, rather than the law. This undermines the case for public provision and public oversight in many contexts. If accountability is understood in instrumental terms as inducing individuals and organizations to act in social welfare maximizing ways, we conclude as follows:

- For-profit firms will generally be more effective at controlling internal agency costs than will public agencies. Agents in for-profit firms face discipline from monitoring by residual claimants, from incentive pay, from the managerial labor market, from hostile takeovers, and from competitive markets and the threat of failure. The threat of failure also serves to discipline private nonprofit firms. Government enterprise, by contrast, has a tenuous link to residual claimants (citizens), is insulated from capital markets (which in turn undermines incentive
pay, the market for corporate control, and the managerial labor market), and rarely fails. These factors imply that governments are less able to motivate their agents than are private firms, whether for-profit or nonprofit.

- The profit motive will in many cases be congruent with social welfare. Since private firms are more efficient, governments should not intervene in these contexts.

- In some cases, an organization’s pursuit of profits is inconsistent with social welfare because of various forms of market failure. But in many of the same cases, and often for the same reasons, government provision will also fail to maximize social welfare. Governments have difficulty establishing incentives for agents both in general and in the particular instances of market failure; sometimes the market failure exacerbates the agency problem in public bodies. Moreover, aside from having inadequate means to accomplish socially desirable ends, government intervention also can result in the pursuit of socially undesirable ends because of the political self-interest of government actors. Only where the cost from market failure exceeds the cost of government failure is intervention appropriate.

- It follows from the preceding propositions that even where some form of government intervention is appropriate, it should rarely entail negating a significant role for the private sector. Intervention should be structured to realize the significant motivational advantages, from profit or from the fear of failure, that arise in the private sector.

- Public accountability mechanisms applied to the private sector, while sometimes appropriate because of the costs of market failure, will often undermine the advantages of private markets. Public oversight is costly and requires reliance on under-motivated and possibly self-interested public agents, which is what privatization is typically designed to avoid.

If, however, public accountability is understood as important in a liberal democracy for non-instrumental reasons, we conclude as follows:

- In many of the domains generally associated with public provision, it is often difficult to discern what those non-instrumental values are. What value does public accountability vindicate in garbage collection, park maintenance, electricity generation, and road repair, for example?

- Proponents of public provision on non-instrumental grounds bear the burden of demonstrating why private accountability mechanisms do not satisfy the non-instrumental values in question, given that private accountability mechanisms are often more effective than those associated with public provision.

- If private accountability mechanisms are in some cases deemed inadequate on non-instrumental grounds, the state can impose a variety of public accountability
mechanisms on private actors. But this strategy is not costless: it risks forfeiting
the benefits of private markets.

Accountability has many dimensions. Accountability mechanisms in the public sector
should not necessarily be the benchmarks against which all accountability mechanisms in
society are measured.

§ 6.3.2. Jody Freeman, *Extending Public Law Norms Through

I. INTRODUCTION

In this Commentary, I suggest a counterintuitive way to view American privatization
trends. Instead of seeing privatization as a means of shrinking government, I imagine it as
a mechanism for expanding government’s reach into realms traditionally thought private.
In other words, privatization can be a means of “publicization,” through which private
actors increasingly commit themselves to traditionally public goals as the price of access
to lucrative opportunities to deliver goods and services that might otherwise be provided
directly by the state. So, rather than compromising democratic norms of accountability,
due process, equality, and rationality—as some critics of privatization fear it will—
privatization might extend these norms to private actors through vehicles such as
budgeting, regulation, and contract.

I say “might” because whether the state’s reach would effectively imbue private entities
with the public norms we attribute to government, or simply constrain them in ways they
would resist and resent, is difficult to predict. As a theoretical matter, however, the
publicization effect seems at least as plausible as the alternative. Surely the state can
exact concessions—in the form of adherence to public norms—in exchange for
contracting out its work. The closest constitutional analogy might be to the federal
government’s authority to induce state conformity to federal goals through its spending
power. Not being constrained by the Tenth Amendment, the federal government has even
greater resources at its disposal to condition and shape the behavior of private contractors
and grantees who receive federal funds to perform arguably public tasks.

State and local governments might similarly impose conditions on their contractors.
Indeed, there is some evidence suggesting that contracting out may extend public norms
to private actors and lead to “re-regulation” rather than deregulation. The publicization
effect would, of course, be highly contingent on a number of factors, including the
political will of Congress and the executive branch, as well as the inclinations of the
judiciary. And the question remains whether this effect can be predictably built into a
privatization scheme and, if so, how best to do it. Specifically, the argument that
publicization is possible must overcome three main objections, which I address in this
Commentary: that publicization will not occur because there is no incentive for

1 pūb'li-ki-zā'shon...
government to demand it; that publicization will not address the most serious risks associated with privatization; and that if publicization does occur, it will effectively turn private actors into public ones, undermining all the benefits of privatization.

The potential for extending public norms to private actors in the way I propose assumes that the features of any privatization scheme are entirely exogenous. That is, there is nothing inevitable about the shape privatization might take: it is an artifact. Indeed, the word “privatization” describes nothing in particular so much as it suggests a host of arrangements. These include, among other things:

(1) the complete or partial sell-off (through asset or share sales) of major public enterprises; (2) the deregulation of a particular industry; (3) the commercialization of a government department; (4) the removal of subsidies to producers; and (5) the assumption by private operators of what were formerly exclusively public services, through, for example, contracting out. Because the last form of privatization mentioned above is the most common of all those used in the United States, and because it remains, at least for some functions, highly controversial, I focus on it exclusively here.

Like every other privatization arrangement on the list, contracting out can take more than one form. On the one hand, government may contract out a function to the market, subject to only minimal contractual commitments, little supervision, and weak regulation. On the other hand, government might insist on detailed contractual terms and on supervising compliance with them, conceivably requiring information disclosure, public consultation, mandatory auditing, and the like. Privatization designed in this second way to maximize accountability might compromise other normative goals one might care about—namely economic efficiency—but this simply highlights the need to make tradeoffs and to strike a balance among competing goals in each case.

This Commentary, which urges a direct focus on such tradeoffs, is part of a broader and longer-term project in which I explore the relationship between administrative law and the role private actors play in public governance. In an earlier article on this theme, I argued that administrative law concepts and doctrines, based, as they are, on a hierarchical and bureaucratic notion of public action, and prioritizing, as they do, constraints on government power, are badly out of synch with reality—one in which many public services and functions are produced by a highly interdependent network of public-private partnerships woven together by history, practice, and mission, and constrained by direct regulation, contract, and informal agreement. The metaphor for this configuration of blended power, I offered, should be contract, not hierarchy. And the unit of analysis at the heart of administrative law, I suggested, should no longer be the public agency as traditionally conceived.

I claimed in that article that administrative law, as currently oriented and constructed, was likely to miss or badly underestimate the potential for harnessing private actors in the service of public goals, or alternatively, that upon recognizing the extent of private involvement in ostensibly public undertakings, administrative law would try to squelch it. I argued that the former would be an unfortunate oversight and the latter a terrible
mistake. I suggested that in addition to conceiving of these public-private relationships as horizontal and negotiated, administrative law scholars ought to rethink the traditional accountability mechanisms they might be tempted to impose upon them. Instead of simply constraining the private role in public governance, I suggested that we aim, instead, to facilitate and direct it. I insisted that private contributions to service provision, and even to regulation, were not necessarily or exclusively dangerous and corrosive of democratic values—which is how they seem to be perceived in mainstream administrative law. I suggested that for-profit firms, and a host of “intermediary institutions,” including nonprofit, professional, religious, and public-interest organizations, can contribute technical innovation, ingenuity, cost-savings, quality, and diversity in the performance of arguably public functions. And I argued that harnessing their potential did not imply a weakened state.

To make my case, I reminded readers that many of the services we associate with government were at one time provided by private actors, and I claimed, with the help of numerous examples, that blended power was now more the norm than the exception. Indeed, this history is well known. What is rather newer, however, is our recognition of this pattern and the perception that existing conceptual categories and understandings of governance need realignment in light of it. The trend toward a more prominent private role also seems to have intensified in recent decades. There are at least two explanations for this intensification: First, as Salamon argues, the “tools” of public action have multiplied and diversified with government expansion. Government’s role in service provision is now often indirect. Second, enthusiasm for greater reliance on market actors and market mechanisms to deliver goods and services has grown in recent decades in the United States, as it has abroad. There is, as a result, a discernible trend toward “privatization,” which in the American context consists largely of contracting out.

This Commentary begins where that earlier article left off. Specifically, it struck me, while reading the literature on privatization, that there remained serious intellectual and practical obstacles to the goal of harnessing private actors in the pursuit of public ends. I noticed, in particular, that economists writing about privatization seem overwhelmingly concerned with cost considerations—and to a lesser extent with quality—to the seeming exclusion of the legal and democratic theory considerations that matter most to public law scholars. Similarly, legal scholars writing about privatization seem relatively unconcerned about the cost and quality considerations that preoccupy economists. This divergence troubled me and suggested the need for bridge-building. Without it, I fear that we will not move to the next, and most challenging, step in the privatization debate: determining when it makes sense to condition private participation in service delivery upon adherence to public law norms, and figuring out how best to implement appropriate terms.

I therefore raise the possibility that privatization might extend public values to private actors to reassure public law scholars that mechanisms exist for structuring public-private partnerships in democracy-enhancing ways. My purpose is to propel the debate beyond the well-entrenched and familiar arguments. To a significant extent, the debate pits economic claims of greater productive efficiency against legal or political arguments that
privatization might compromise democratic norms. I describe the status of this debate in Part II.

Unfortunately, this opposition of economic and noneconomic values suggests that privatization represents a zero-sum game between public norms and private power. This oversimplifies the choices presented by any privatization decision. It erroneously suggests that we must sacrifice one set of goals entirely to the other because the two sets are fundamentally incompatible. I maintain that we might find ways to structure privatization that allow us to have some of each. To this end, I argue in Part III that contracting out could lead to publicization under some circumstances and that publicization need not undermine all of the advantages of privatization.

I do not mean to suggest that we can have everything at once, however. Because it will be costly and burdensome to extend public norms to private actors, not every case warrants this treatment. Nevertheless, despite the appearance of wide divergence between the economic and public law views of privatization, there is, perhaps surprisingly, significant convergence on when privatization makes good sense and when it seems risky. There might be broader agreement than we think on three questions: when we should privatize, when we should not, and when we should pursue publicization rather than give up on privatization altogether.

In Part IV, I propose three considerations that, from the public law perspective, distinguish stronger cases for publicization from weaker ones. While I express these factors in the noneconomic vocabulary of the public law perspective, many of them have parallels in the economic view. These factors are (1) the relative precision with which a service can be specified and the extent of the provider’s discretion, (2) the potential impact on the consumer, and (3) the government’s motivation for privatization. The argument for publicization is strongest in instances when services are highly contentious, value-laden, and hard to specify, and when providers enjoy significant discretion; when services affect vulnerable populations with few exit options and little political clout; and/or when the motivation for privatization is explicitly ideological or clearly corrupt. These are not the exclusive factors to consider, but they reflect the most pressing concerns of the public law perspective on privatization, and this perspective must be given voice in a way that will facilitate engagement with adherents of the economic view. In identifying these factors, I mean simply to begin a conversation about when we might condition privatization on adherence to democratic norms. I invite others to identify and defend alternatives.

II. The State of the Debate

A. Ideological Arguments

The publicization effect runs directly counter to the intentions of the most ideologically motivated privatization proponents, who seek to shrink the state in the name of individual liberty. Yet it might cohere very comfortably with the goals of “pragmatic privatizers” who wish to improve quality and cut the costs of service provision consistently with democratic values. Understanding why publicization is consistent with pragmatic goals,
however, requires a word of explanation about the difference between the two approaches.

In the last two decades, privatization has been championed by conservative policymakers, academics, and public intellectuals as instrumental to reducing the size of government and broadly restructuring society in line with a conservative agenda. Privatization coincides with other political and economic developments—including globalization, free trade, market integration, and deregulation—that similarly reinforce an ideological preference for private over public ordering and market over noneconomic values.

This preference for market ordering and market values is part of a broader cultural and political shift. At least in the United States, the last few decades have been marked by an increased faith in markets and a corresponding decrease in support for public institutions. Over this period, domestic politics has reflected, if not contributed to, this waning enthusiasm for government. Ronald Reagan’s presidency, which stressed deregulation, embodied this antigovernment ethos. Nor has it been limited to Republican administrations; former President Bill Clinton, a self-identified “New Democrat,” declared the era of big government over, ended “welfare as we know it,” and instructed former Vice President Al Gore to conduct a “performance review” of government agencies to create a government that “works better and costs less.” In so doing, the Clinton Administration adopted the rhetoric and conceptual approach of the “New Public Management,” which proposes to organize public bureaucracies more like private firms. A distinctive feature of this approach is its view that taxpayers are “customers” interested primarily in efficiency. At the same time, the Administration shrank the federal civil service to a significant extent, relying on a host of private contractors to perform a wide variety of functions and provide a broad range of services.

The popular preference for the market over government has continued with the election of President George W. Bush, who ran for office on a largely antigovernment platform (including an initiative to fund private faith-based institutions to deliver social services) and whose ideological commitments in this regard seem much stronger than his Democratic predecessor’s. Bush campaigned as a CEO who would run government as a business. It is not surprising, therefore, that even as he is now creating a new Department of Homeland Security, mobilizing for war with Iraq, and bolstering law enforcement and security-related agencies in response to the events of September 11, 2001, he has also stuck to his small-government guns by seeking to contract out nearly half the federal civilian workforce.

Ideological support for privatization is matched by a contrary ideological view that holds the public sector in higher esteem. Adherents of this view are more likely to regard government as a force for progressive ideals. They would rule out privatization as a presumptive matter, at least for a subset of functions that go to the heart of what they see as the state’s inherent responsibilities in a liberal democratic society. These functions might include criminal and civil adjudication, policing, incarceration, education, transportation, health care, national defense, and foreign policy. Like the pro-privatization position aimed at limiting government, this anti-privatization rejoinder is grounded in political philosophy rather than sound economics. It holds that even if
privatization were proven to deliver significant cost savings, it would still be an illegitimate choice for organizing service provision, at least for those functions that are so inherently governmental as to be categorically nondelegable.

The trouble with ideological positions of this sort, on both sides of the privatization debate, is their intransigence in light of arguments about the empirical consequences of privatization. Empirically grounded claims about how privatization works, or might be made to work, are unlikely to move either side from its core beliefs. For this reason, the possibility that public values might extend to private actors through privatization would be less interesting to ideological privatizers than to pragmatic privatizers. Pragmatic privatization arguments, though not free of ideological underpinnings, are potentially more responsive to arguments about the benefits of privatization and the potential for structuring privatization to ensure that it does not erode democratic norms.

B. Pragmatic Arguments

A variety of arguments about privatization seem more pragmatic than ideological. As an illustration, I have in mind the paradigmatic public manager—a county official, say—faced with the task of providing high-quality services in a cost-effective way. Pragmatic privatizers of this sort seek more efficient service delivery while adhering to numerous and sometimes conflicting mandates from higher levels of government. Operating under considerable fiscal, institutional, and legal constraints, these local officials try to solve immediate budgetary and service-provision problems rather than seek to transform society’s relationship to government.

So, when a local agency contracts out highway maintenance or garbage collection primarily because of cost considerations, that decision may qualify as an instance of pragmatic privatization. While such choices have political implications (such as systematically shrinking the civil service), they may not be primarily motivated by political expediency or ideology.

From this pragmatic perspective, privatization is a means of improving productive efficiency: obtaining high-quality services at the lowest possible cost, thereby freeing up resources that might otherwise go to waste and allocating them elsewhere to maximize welfare. Such pragmatic arguments sound largely in the discourse of economics and tend to elevate economic values over other goals.

However, maximizing efficiency in service provision implicates two separate questions: First, who should set the level at which the relevant good is provided and dictate the terms of its provision? Second, and independently, who should provide it? Pragmatic privatizers typically focus on the second question, taking for granted that government has a responsibility to establish service levels but viewing government as ill-designed and poorly equipped in most cases to deliver services directly.

This view that private entities are better suited to deliver most goods and services reflects a comparison of public and private organizations in which public institutions fare badly. Economists point out in this regard that civil service protections reduce the employer’s
flexibility to reward and punish employee performance, making it harder to encourage employees to produce high-quality services and cut costs. Public agencies are generally not client-oriented, economists argue, so agency employees lack incentives to perform well. Agencies are not profit-maximizing entities disciplined by competition and the possibility of bankruptcy, and so will tend to be both less innovative and more wasteful. Because of differences of structure, organization, and institutional culture, private firms are thought to be capable of providing the same or higher-quality services at lower cost than can public agencies.

The stakeholder at the heart of this economic perspective is the taxpayer-consumer on whose behalf the privatizer pursues the twin goals of lower cost and greater technical efficiency in providing the service. To the pragmatic privatizer, it matters little whether the service under consideration is waste collection, power generation, education, or incarceration. Similarly irrelevant are the vulnerability of the population being served, its exit options, and its political power. Any of these services might be ripe for privatization if they present opportunities to cut costs and improve service quality through innovation. By contrast, these other, noneconomic considerations might give pause to others who approach privatization with a different set of priorities.

A word of caution: I do not wish to conflate pragmatic arguments for privatization as if there were only one economic approach to the matter. The economics literature on privatization is voluminous, diverse, and nuanced. But it is fair to say that pragmatic arguments typically draw on economic conceptions of the advantages of private over public service provision and not on other normative frameworks. And the bulk of these studies focus on cost (for example, monitoring and information costs) rather than on quality.

Importantly, adopting a pragmatic approach to privatization is not the same thing as being pro-privatization. Critics may argue against privatization by invoking the same kinds of economic considerations listed above. They might argue that the anticipated benefits of privatization depend on market conditions that frequently fail to materialize. Privatization efforts may fail because of an absence of competition (in some cases, privatization simply replaces a public monopoly with a private one) or because of political obstacles that hamper the efforts from the start. Those who object to privatization for such reasons have no ideological bone to pick (or, if they do, it is secondary); they simply doubt that privatization will deliver equal or greater quality at lower cost. If economic conditions were different, they might enthusiastically support privatization.

Still, such arguments against privatization are relatively few, or limited to activities where privatization seems unfathomable (such as foreign policy or national defense). The economics literature is generally optimistic about privatization for most other services and tends to adopt the view that privatization should be the default assumption for service delivery unless a convincing economic case can be made against it. This optimism persists even in the midst of widespread acknowledgment that the presence of some conditions (for example, high asset specificity and noncontract[ible quality]) and the absence of others (for example, competition) might reduce privatization’s prospects for success.
As we shall see in the next section, however, there is another, more skeptical approach to privatization that prioritizes legal and democratic values. This latter view tends to counsel against privatization. Together, these two perspectives currently frame the debate as a choice between incompatible normative systems, one grounded in economics with its focus on cost and, to a lesser extent, quality, and the other in legal and democratic theory with its focus on rights and accountability. These two perspectives suggest that privatization is a zero-sum game between public and private power. To choose privatization is to enhance the private sector at the expense of government. As I argue in Part III, both of these impressions are misleading.

C. The Public Law Perspective

Another species of argument sometimes invoked to oppose privatization is legal rather than economic. It, too, is pragmatic, in the sense that it is concerned with the consequences of privatization, but it is also ideological in its baseline commitment to democratic norms. Proponents of this perspective might concede that productive efficiency in service provision has value, and they might be in favor of cost savings in service provision generally. But they would also worry about privatization’s implications for what they regard as liberal democratic norms of accountability, due process, equality, rationality, and the like. Adherents of this perspective could never be called non-ideological—they would, for example, reject out of hand the radical conservative agenda of the pro-privatization ideologues—but they might tolerate a somewhat smaller and reconfigured government in exchange for improvements and cost savings in public services. However, while they might be willing to debate the appropriate size, structure, and role of government, they would argue (as Martha Minow does in this volume) that whatever the ultimate public/private mix, privatized service provision must be accountable and consistent with democratic norms.

For convenience, I will call this perspective the “public law” view because I associate it with public law scholars, whose central concern is the relationship between government and private citizens. To my mind, the public law view is represented paradigmatically by administrative law scholars for whom the substantive legality and procedural regularity of government action is the primary preoccupation—though it is certainly not limited to this group. Adherents of the public law perspective begin with the observation that the Constitution, together with statutes like the federal Administrative Procedure Act (APA), imposes on government a host of obligations designed to render decisionmaking open, accountable, rational, and fair. While they recognize that such impositions generally apply only to government, they bristle at the notion that government would seek to avoid these obligations by systematically contracting out functions.

As a result, the public law perspective asks not whether privatization is efficient, but whether it erodes the public law norms that these constitutional and statutory limits are designed to protect. In particular, this perspective prioritizes legally required procedures designed to guarantee public participation and due process, such as those required by the notice-and-comment and formal adjudication procedures of the APA. Compliance with these procedures has an inherent value in the public law view; it is part of the minimum obligation a state owes to its citizens. Procedural regularity is also instrumental,
however—procedures designed to ensure public participation and individual fairness might improve the rationality of decisionmaking and legitimize the authority of the state.

To round out the comparison with the economic perspective described above, the relevant stakeholder in the public law perspective is the diffuse public, to which public law scholars ascribe an interest in such things as accountability, due process, and rationality in decisionmaking. They do this as a normative rather than an empirical matter, believing a commitment to such standards fundamental to a liberal democracy. In this same vein, public law scholars have traditionally sought to ensure that government engage in policymaking in the “public interest”—a notoriously ill-defined term, but one frequently invoked to convey a preference for deliberative, disinterested, and expert decisionmaking that does not merely serve the interests of a special few.

Public law scholars worry that privatization may enable government to avoid its traditional legal obligations, leading to an erosion of public law norms and a systematic failure of public accountability. How might this happen? First, Congress has wide discretion to delegate functions to both public and private actors. The nondelegation doctrine does little to limit Congress’s prerogative in this regard, provided that Congress supplies the requisite “intelligible principle” to guide the exercise of private discretion. State-action doctrine insulates most private actors from constitutional obligations, including due process, that would apply to the government if it performed the same functions. Though the Supreme Court occasionally treats private actors as state actors for constitutional purposes, these instances are few.

Beyond constitutional limitations lie statutory imperatives designed to promote public law norms. Private actors may effectively escape these as well. First, the APA specifically excludes both grants and contracts from the demands of notice-and-comment rulemaking normally applicable to federal agencies. Moreover, the APA subjects only agency action to potential judicial review. As part of their inherent power, agencies may further delegate powers entrusted to them to other actors with virtually no constitutional limitation, provided that Congress does not prohibit them from doing so. While an agency’s decision to contract with private providers may be subject to procurement regulations if the good or service purchased is for government consumption, the implementation of federal grants and run-of-the-mill discretionary funding decisions are not governed by the procurement process and generally receive considerable judicial deference. Courts consider decisions to award discretionary grants to be informal policymaking, a mode of agency action over which courts hesitate to tread too heavily. Moreover, courts have referred to federal grants as “gifts.”

The bulk of privatization also remains beyond the reach of the subconstitutional discretion-constraining and accountability-forcing mechanisms of administrative law. Private actors need not comply with any of the APA’s procedural requirements, nor must they, generally speaking, observe the disclosure provisions of the Freedom of Information Act (FOIA) or the open-meeting obligations of other sunshine laws that apply exclusively to government actors. Private contractors notably escape the civil-service protection rules, as well as the conflict-of-interest and ethics rules that apply to
public employers. Simply receiving a government grant does not turn the recipient into a
government agent.

To illustrate how privatization can raise public law concerns in both obvious and subtle
ways, consider welfare reform. The elimination of Aid to Families with Dependent
Children and its replacement with Temporary Assistance for Needy Families (TANF) via
the Personal Responsibility and Work Opportunity Reconciliation Act of 1996
(PRWORA) created a new welfare policy in which the federal government replaced a
federal entitlement program with block grants to the states. Among the block grant
conditional terms, PRWORA requires states to ensure that recipients comply with work
specifications and time limitations. PRWORA devolves policy control to states and,
ultimately, to local governments, to which states generally assign program authority.
These local governments contract with private providers to furnish welfare-related
services.

Critics claim that this new, highly decentralized approach has resulted in a transition from
a bureaucratic and rule-oriented model of welfare to a discretionary and market-based
one. The role of public agencies has been refashioned through the adoption of market-
based performance measures and other private-sector techniques of “entrepreneurial
government.” As Diller puts it, the public-benefits agency now sees itself as a corporation
selling a product: self-sufficiency.

Devolution has also been accompanied by a trend toward privatization. Although many
welfare services have long been provided by a network of local agencies and private
organizations, under the new TANF program, administrative functions once reserved to
public agencies are increasingly contracted out. Many states have privatized such
functions as case management, child care, transportation, education, job training, and
placement services for TANF recipients, and at least one state had sought to put its entire
program, including eligibility determinations, out for bidding to private firms before the
Clinton administration prohibited the outsourcing of some eligibility determinations.
Moreover, the Act’s new emphasis on “welfare to work” and its clear intention to change
the “institutional culture” of welfare administration encourage reliance on private, for-
profit firms.

As a result, critics claim, public law protections have eroded. To some extent, procedural
protections for welfare recipients have diminished as a result of devolving discretion to
ground-level workers generally, but privatization has exacerbated the new accountability
problems. For example, contracts with private providers are typically awarded without
notice and comment, preventing TANF recipients and their advocates from influencing
the contractual terms or, at a minimum, having their views heard in the process of
constructing the contracts. The new array of contracts has produced a more fragmented
network of local agencies and private providers, which makes it more difficult to obtain
information, lodge complaints, and monitor quality. And fewer welfare-related
decisions—many now made by private contractors—are considered “determinations” or
“denials of benefits” which, if made by an employee of a public agency under the old
regime, would be subject to due process hearings and judicial review.
There are at least two responses to such concerns about privatization. First, all of the accountability-forcing mechanisms described above were originally designed to constrain state power and its potential to intrude on individual rights. One might, therefore, choose not to worry about privatization on the grounds that private power is simply less dangerous than state power; indeed, this is the fundamental principle reflected in the state-action doctrine. In view of this argument, the only legally relevant question about government contracting out is whether it is constitutionally permissible and otherwise lawful under applicable statutes. If the answer is yes, as it certainly is in the context of the new welfare law, that ought to end the inquiry from a legal perspective.

This, however, is a formalistic way to dispense with the public law concern about the erosion of public law norms in a world where multinational corporations and international organizations wield enormous power. Though government is still uniquely powerful because of its monopoly on state-sanctioned force, whether a function or service is dangerous does not turn solely on whether it is publicly or privately provided; in at least some instances, the nature of the provider is far less important than the nature of the service and how it affects consumers. For example, the risks to consumer safety and well-being as a result of private monopoly control over utilities such as gas and electricity led courts in the nineteenth century to treat these enterprises as businesses “affected with a public interest” and to impose minimal due process norms on them. In any event, it is the transition from government to private provision that the public law perspective finds problematic, especially when considered cumulatively, because it allows government to avoid obligations it would undertake if it provided services directly.

Second, one might argue that contracting out is inconsequential because government typically sets the level of service provision itself and delegates only implementation to private providers. This arrangement preserves traditional accountability while shifting responsibility for delivering the service to the private contractor, a distinction that, one might argue, critics tend to confuse. Again, this claim is too formalistic to provide much comfort to those who share the public law perspective. Providers (whether public or private) enjoy considerable discretion when implementing the policy choices of elected officials. This discretion affords them an opportunity to redefine policy choices or to specify them at a level of detail unanticipated by policymakers. Decisions at the ground level of policy implementation can be as consequential as decisions, such as eligibility standards or general program directives, set directly by a centralized public authority. Even the power simply to provide the most carefully specified services creates a principal-agent problem in which a contractor may, without violating any technical contractual terms, enjoy substantial room to maneuver.

Because the public law perspective so values the substantive and procedural protections embodied in the Constitution and statutes like the APA, extensive contracting out of service provision under conditions that effectively insulate private actors from scrutiny poses risks that many adherents to this view are not willing to take. And while they might acknowledge that privatization can cut costs and improve quality in some circumstances, adherents of the public law view tend to think cost and quality considerations should not be decisive when these legal values, which ensure public participation, accountability,
and due process, might be sacrificed. This is especially true with the provision of benefits and services that affect vulnerable populations lacking exit options, especially where important individual rights and interests are at stake.

D. The Impasse

It is easy to see how the pragmatic debate over privatization can devolve into a conflict between economic arguments in favor of privatization and public law arguments against it. For pragmatic privaters focused on maximizing the economic benefits of private service provision, the public law appetite for expensive and time-consuming processes seems counterproductive and dismissive of budgetary realities. Procedures like notice-and-comment rulemaking, administrative hearings, and even judicial review are part of the problem privatization is trying to solve. This sentiment might arise for at least two reasons: first, because procedures can be costly to implement, and second, because they might be seen as performing a political function—enfranchising interest groups to participate in agency implementation—that is economically wasteful. Indeed, one of the purposes of privatization is to replace a rule-oriented, bureaucratic system of accountability with the more streamlined accountability afforded by the market.

Part of the gulf between the two perspectives may be due to the compartmentalized and self-referential nature of the privatization literature: most analyses seem oriented to either one community or the other and aimed at those who share common disciplinary tools and normative frameworks. But perhaps one can express the value of procedure in terms that pragmatic privaters would appreciate. One might argue, for example, that notice-and-comment rulemaking feeds information into the administrative decisionmaking process at the implementation stage, when agencies translate legislative generalities into operational specificity. In this way, legal procedures provide a mechanism for preference elicitation and preference aggregation, which economists might consider valuable.

Broad consultation with interested parties who have a concrete stake in an issue can also force the agency to consider a wide range of scientific and economic data it might otherwise ignore or treat less seriously, which can help minimize costly policy errors. While adherence to procedure alone can never guarantee rational decisionmaking, it might plausibly improve rationality, especially when the decisionmaker must account for policy choices to a court bound to scrutinize those decisions for arbitrariness. Finally, legal procedures and judicial review provide a number of other hard-to-quantify benefits, including a minimal guarantee of fairness and the political legitimacy that comes from the perception of fairness.

Just as proponents of the economic perspective might come to acknowledge the value of legal procedures, public law scholars might learn to appreciate the economic benefits of privatization. The public demands—and government has undertaken to provide—a broad range of services on a limited budget. Pragmatic privaters might seem like bloodless cost-cutters from the public law perspective, but the purpose of shifting provision to the private sector is to minimize costs and ultimately better allocate expenditures to maximize welfare. Pragmatic privaters are always thinking of opportunity costs in the context of budget constraints: to spend more on one service, like road repair, is to spend
less on something else, like job training or daycare. While public law scholars are aware of opportunity costs, concern about such costs does not feature prominently in their critiques of privatization.

Notably, pragmatic privatizers also seek to improve quality of service in addition to reducing costs. Especially where gains from innovation might be high (for example, in health care and education), they wish to capitalize on the capacity of the market to develop new approaches. It is hard to quibble with this goal.

Consider again welfare reform. Though the PRWORA has attracted almost universal scorn from the welfare advocacy community, there is a plausible story to tell about how increased contracting with private providers may actually, under the right circumstances, diversify the kind of services that welfare recipients need most—including child support and child care, job training and placement, transportation, and emergency services—and improve quality in addition to saving costs. This is not pie-in-the-sky idealism; it is a matter of design and political will.

As Roper notes in his study of New Jersey’s implementation of the PRWORA and the state’s own welfare reform plan, the most recent iteration of welfare reform is not responsible for the emergence of private social-service providers. Indeed, welfare-related services have been provided by local and primarily nonprofit organizations since the 1970s. Roper’s report does confirm that the array of local organizations has expanded in recent years beyond churches, community development corporations, and advocacy groups, to include more for-profit firms. And reflecting nationwide trends, New Jersey has increasingly adopted a pro-privatization attitude and pushed to increase contracting with private for-profit firms. Specifically, the Whitman administration sought to implement three reforms: performance-based contracting (assessing outputs, not inputs), service-provider diversification, and service integration (reducing service duplication and consolidating services).

These are laudable goals. Performance-based contracting could introduce greater accountability into a system in which providers have historically been evaluated based on the number of people served rather than on the number of people placed in jobs. The new emphasis on job placement can itself create perverse incentives (for example, placing people in unsuitable and short-term jobs or “creaming”—serving only the most qualified among those in the eligible pool), but at least this approach seems to be on the right track by focusing on outcomes and trying to measure them.

In practical terms, diversifying the number of provider organizations is critical to the weakening of long-term relationships between existing providers and powerful local contracting agencies. These relationships tend to keep existing contracts in place regardless of the quality of performance, and they enable existing providers to obstruct managerial and programmatic reforms that would reduce the flow of public funding to them. So while provider diversification might have an explicitly political motivation, that motivation could result in less patronage and corruption, both reducing costs and improving quality. Diversification can also help introduce new perspectives into service provision. New Jersey explicitly sought to enlist local minority communities, including
faith-based organizations as well as neighborhood and community groups, in the
provision of social services.

And finally, it is hard to oppose the efficiency goals of service integration, which favors
“bundling” services to achieve economies of scale and ideally results in more “one-stop
shopping” for those needing services.

Without denying the legitimacy of Diller’s and others’ critique of the apparent
ideological and political motivations behind welfare reform, the move to privatize, at
least as a theoretical matter, is defensible. At a minimum, one must take the pragmatic
impulses behind such reforms seriously.

Indeed, pragmatic supporters of privatization have a point about the need to do
something. Many publicly provided services deservedly attract criticism for being costly
and inadequate . . . . One need not be a conservative ideologue to criticize the quality of
education in public schools, the dangerous conditions in overcrowded prisons, and the
dilapidated state of much urban infrastructure. If private provision might improve these
services and do so for less, surely it is worth a try. Feelings like these explain why
parents of all political persuasions might support school vouchers, for example, even at
the risk of having their children attend private schools where they would be exposed to
religious instruction the parents would rather avoid. Whatever its downside, these parents
might think, privatization has to be better than the status quo.

Where does this debate between the economic and public law perspectives leave us?
Good translation across the disciplinary, conceptual, and linguistic divides between the
two views only goes so far. Even if the prospect of cost savings and improved quality
seems promising, many adherents of the public law view would decline to take the risks
that, in their estimation, widespread contracting out entails. Likewise, although
economists might acknowledge the value of legal procedures and public law norms, they
may devalue them relative to other things; they may even lack a ready mechanism for
building such norms into their models. And each side, might, to some extent, resist the
normative priorities of the other. To an outside observer, the choice can appear to be all
or nothing: either privatize and hope for the best, or refuse to privatize for fear of the
worst.

Any suggestion that the conflict between the two perspectives is intractable in every case,
however, is questionable. Perhaps there is a way to take advantage of the strengths of
each without ignoring the legitimate claims made by the other. I turn to this possibility in
Part III, below.

III. PRIVATIZATION’S POTENTIAL TO EXTEND PUBLIC LAW NORMS TO PRIVATE ACTORS

One way to mollify adherents of the public law perspective would be to argue that
privatization could in fact extend public law norms to private actors under some
circumstances—a process I call publicization. Perhaps one can obtain some of the
benefits of privatization without systematically eroding democratic norms. Put most
ambitiously, perhaps privatization can enhance public law norms by extending them to

Draft — do not distribute!
realms where they typically do not play a significant role. In this vision, public and private power might grow simultaneously, services might improve and diversify, and the worst fears of the public law perspective need never materialize.

Such a claim will no doubt be met with skepticism, and I wrestle later in this Part with what I anticipate to be the most significant objections to it: that there is no incentive for government to provide for publicization, that it will not work, and that if adopted, it would effectively turn private providers into public agencies, thereby sacrificing the benefits of privatization. Before defending my proposal, however, I consider below, as a thought experiment, how publicization might come about.

A. Legislative Mechanisms

First, the mechanisms to accomplish publicization exist and are by no means new. Since the late nineteenth century, federal and state governments have imposed requirements on both private for-profit corporations and nonprofit organizations through direct regulation, conditioned grants, and the conferral of tax subsidies, among other instruments. Common law courts have long imposed legal restrictions on private actors through evolving tort, contract, and public nuisance standards. Extending compliance with public law norms to private actors engaged in service delivery would, therefore, be an expansion of a traditional practice.

To illustrate, Congress could extend by statute some or all of the public law requirements discussed in Part II to private actors—requiring them to disclose information in a public record, hold public hearings and take public comment, and justify their programmatic decisions with reasons that would be subject to judicial review. Congress could amend the APA to eliminate the section 553 exemption for contracts and grants. This would let members of the public participate in agency decisions related to these instruments, which would help to produce a record for judicial review. Congress could also extend other statutes, such as FOIA, to private entities.

Congress could accomplish much of this indirectly, through the federal grant system pursuant to its spending power. Indeed, Congress already conditions federal grants extensively, demanding that states adhere to federal program goals and other restrictions, and requiring state and local governments to regulate private contractors in a variety of intrusive ways. The grant device lets Congress in effect simultaneously purchase services provided by third parties for the benefit of other third parties and regulate the delivery of these services through specific legislative requirements, the implementation of which is delegated, typically, to a federal agency. For example, when a state wishes to receive federal Medicaid funding, it must file a plan that assures compliance with various provisions of the Social Security Act, the Balanced Budget Act, and a host of federal regulations promulgated by the Department of Health and Human Services (HHS). Among other things, federal regulations include requirements that apply to contracts between the state agency and the private institutions that deliver health care, such as hospitals and managed care organizations.
As a condition of federal grants, Congress could, therefore, tailor privatization experiments to extend federal goals not only to the state and local government grantees that directly receive the funds, but also to private contractors charged with service delivery. Congress might minimize the discretion of private contractors by specifying performance criteria or dictating substantive contractual terms, including requirements for regular and detailed reporting. Congress might demand closer or more extensive monitoring of private contractors by federal agencies, not only for cost control and fraud prevention purposes, but also for quality control, which these agencies charged with oversight have traditionally not done very effectively. And to accomplish this, Congress might fund and empower agencies to monitor in a serious way. This is something that agencies—already frequently underfunded to carry out their basic regulatory and adjudicative tasks—are generally ill-equipped to do. In addition, Congress might provide for supplementary contract enforcement by granting private rights of action in legislation, or by requiring contracts to provide for suits by third-party beneficiaries.

Federal agencies charged with awarding grants could themselves, in many cases and without any additional delegated authority, impose a wide variety of substantive regulations on recipients of those grants or on contractors with whom the grantees do business. Despite legislative prescriptions and the coordinating directives of the Office of Management and Budget, agencies generally have considerable discretion over grant implementation. And finally, even without congressional or federal agency action, state governments could adopt similar measures independently, either through direct regulation of private contractors or by inserting terms into the contracts themselves.

B. Judicial Mechanisms

While courts have thus far largely declined to curtail the trend toward privatization, they might increasingly scrutinize both legislative and bureaucratic efforts to contract out government functions for potential constitutional or statutory violations. Invalidating instances of contracting out would not extend public norms to private actors, so I do not focus on them at length here. Suffice it to say that the Supreme Court may find some functions nondelegable. The Court has already suggested that in addition to conducting elections and running towns, some other government functions may be nondelegable, such as tax collection, fire and police protection, and education. Perhaps as experience with privatization develops, the Court will expand this list. Courts could, moreover, conceivably invalidate private delegations while leaving delegations to public agencies in place. Jack Beermann has suggested, along these lines, that the Supreme Court’s concern with political accountability, evinced in its Tenth Amendment jurisprudence, could provide the basis for invalidating some delegations to private actors on grounds that they will confuse the public about whom to blame when things go wrong. In addition, many state constitutions contain “accountability clauses” that may limit privatization.

Short of invalidating contracts with private providers outright, however, courts may choose to scrutinize those contracts more carefully to ensure they provide for adequate supervision by Congress. Alternatively, the Court might soften its approach to state action, perhaps relaxing the conditions under which it is willing to find that private contractors are performing functions traditionally and exclusively performed by the state.
This move would limit the impact of broad delegations and provide a mechanism for extending public norms to private actors without invoking the blunt instrument of the nondelegation doctrine, which the Court so resoundingly and unanimously declared defunct in the recent *American Trucking* case.

The Supreme Court has spoken on at least two occasions about privately run prisons or detention centers, with the result that these are now equally and, some might argue, more accountable as a constitutional matter than public prisons. While private prisons may still present dangers that public prisons do not, the Court has shown some willingness to ensure that privatization does not compromise constitutional norms, at least with regard to a sensitive function such as incarceration. This is just one example, but it offers some reason to believe that courts might fill the gap should legislatures and agencies fail to extend public norms to private actors exercising their contractual duties.

And while courts have been reluctant to construe statutes to afford plaintiffs private rights of action absent express congressional intent, courts might construe contracts to afford third-party beneficiaries the right to sue to enforce contractual terms, whether or not Congress or state legislatures expressly build such causes of action into contractual schemes.

With regard to agency decisions to issue grants and award contracts, courts might simply determine that the relevant governing statutes do not preclude judicial review and then impose on agencies a duty to oversee their contracts more vigorously under pain of being found to be acting arbitrarily or capriciously. In other words, courts might jettison the traditional deference they afford such informal policymaking if it becomes the primary tool of contracting out traditional government functions and if courts detect a serious erosion of public law norms. Critics of regulatory reform have called for a similar reconsideration of judicial deference to an agency’s exercise of enforcement discretion, when that discretion has been arguably used to undermine otherwise mandatory statutory or regulatory requirements. Courts are free to scrutinize agency contracts and grants for arbitrariness, even when those contracts are not subject to the APA’s notice-and-comment requirements for rulemaking, unless Congress provides otherwise.

Short of such steps, courts might simply elaborate common law contract principles that require fair procedure and rational decisionmaking to force private actors to comply with rudimentary due process. While these cases have historically applied to health care providers excluded or expelled from private associations on which they depended for their livelihood, there are examples of such obligations applying more broadly. Recently, a California court imposed procedural obligations on a private employer seeking to dismiss an employee for sexual harassment. It is conceivable, then, that courts might find such obligations applicable to private contractors in control of an important service when they are in contractual relationships with consumers (for example, nursing home residents and utility customers). The California Supreme Court has held, for example, that private utilities are bound to comply with antidiscrimination norms because of state-conferred monopoly status and pervasive state regulation. In addition, private actors are generally more vulnerable to tort liability than public entities. It is conceivable that courts will increasingly rely on tort law as a mechanism for disciplining private decisionmaking.
In addition to constitutional, administrative, contract, and tort law doctrines, corporate law may be a driver of publicization, primarily through information disclosure. Beermann argues that corporate law functions much like administrative law, to force accountability in private decisionmaking. For example, corporations must disclose information and submit to public scrutiny, much as public agencies do. The securities laws and the general demands of capital markets “place a great deal of information about private corporations in the public domain. With regard to publicly traded corporations, it is not hyperbole to state that nearly as much information about them is easily accessible to the public as is available about the workings of many administrative agencies.” The investment and insurance industries are both subject to disclosure requirements, either as a result of legal requirements or because consumers demand information.

Finally, Beermann argues, litigation also produces information about private companies. He likens liberal discovery rules to the FOIA and views judicial deference to corporate decisions under the “business judgment rule” to be roughly parallel to judicial review of agency decisionmaking under the arbitrary or capricious standard.

Beermann claims that government, beyond requiring information disclosure, already tends to regulate contractors more stringently than other private entities; he cites the federal government’s extension of nondiscrimination norms to private contractors via executive order, long before the Civil Rights Act imposed legal requirements to this effect more broadly on private entities. As a result of the greater demands placed on private contractors, privatization can, in some instances, increase political accountability—defined as “amenability of a government policy or activity to monitoring through the political process”—rather than compromise it. Charter schools offer another example of how regulation constrains some forms of privatization. Charter schools may be more heavily regulated than public schools because they are chartered by the state and regulated by state agencies, rather than by local school boards, and charter schools can be abolished for failing to meet government-established performance standards.

Beermann may be somewhat optimistic about the extent to which corporate law doctrines currently discipline private firms. For example, information disclosure alone may do little to make private firms accountable in the absence of an oversight body equipped to discipline them for infractions. (The recent spate of accounting fraud at Enron, Tyco, and other corporations comes to mind.) Moreover, the analogy between corporate disclosure and agency accountability seems overdrawn because agencies are accountable to many different overseers, including all three branches of government. In addition, some constitutional law scholars will no doubt find Beermann’s proposal about the extension of Tenth Amendment norms to the privatization context unrealistic.

Moreover, the extent to which public norms already extend to private actors may not be as great as Beermann’s civil rights example suggests. After all, as noted earlier, and as Beermann concedes, the disclosure and open-meeting laws that apply to government still, as a rule, do not apply to private organizations. And finally, charter schools are arguably the most constrained example of all the educational reforms proposed by the “school choice” movement. Vouchers are the most extensive form of privatization in this context, and the regulations tied to voucher money appear to be significantly weaker, though they
could conceivably be strengthened. As Justice Breyer pointed out in his dissent in *Zelman v. Simmons-Harris*, as a condition of accepting public vouchers, private religious schools could be bound to comply with a variety of intrusive requirements, including antidiscrimination norms.

Still, Beermann’s argument serves as a reminder that regulating private actors as a condition of awarding them contracts is nothing new. Private firms and nonprofit organizations—including charities—are already forced to adhere to public law norms to some extent, as a consequence of either direct regulation or licensing, or as a condition of receiving government grants or tax-exempt status. The federal government already uses its spending power to condition federal funding on state and private compliance with antidiscrimination norms. And already, some federal government programs extend nondiscrimination requirements beyond state and local government grantees to private subcontractors.

The British privatization experience, while specific in some ways to its context, offers an instructive illustration of how contracting out can extend traditional public law norms to private providers. Prosser recounts how the privatization of British utilities such as gas, rail, telecommunications, and electricity led to the unexpected emergence of a new “public service law” consisting of consumer protection regulations applicable to private providers. This public service law consists of a set of obligations to “make basic public services available to all citizens without discrimination,” including such consumer protection measures as ensuring universal telephone service, protecting gas and electricity customers from disconnection, and prohibiting “cream skimming”—offering service only to the most profitable customers.

At the time of privatization, Prosser insists, the British government had no intention of burdening newly privatized utilities (transformed in the British case from nationalized industries) with legal obligations. But obligations emerged. Initially, they were authorized in fairly vague terms by the legislation that established the new regulatory bodies—one independent regulator for each utility—charged with overseeing privatization. Ultimately, however, these obligations took concrete shape once the independent regulators imposed burdens on private providers, stepping beyond their main task of acting as specialist competition authorities. The principal vehicle of imposing these restrictions on private decisionmaking has been licensing.

While licensing as a form of regulation is hardly new, especially to those familiar with the American regulatory system, its use to extend public law norms of universal access and nondiscrimination to freshly privatized services in the context of Britain’s ideologically motivated privatization thrust was, to say the least, surprising. As Prosser points out: “The creation of independent regulators was originally viewed as a way of increasing investor confidence by preventing arbitrary intervention by governments after privatization. Paradoxically, however, it has formed the most important basis for the creation of an emerging body of public service law.” Notably, these layers of regulation were imposed under a Conservative government and have only intensified since.
In invoking this example, I mean only to suggest that such a turn of events is possible, not inevitable. British privatization, which consists to a much greater extent of selling state assets, does not provide a mirror for the American experience. But in either system, once the choice is made to rely on private service provision, the state may still exercise residual regulatory control. To be effective, licensing arrangements like those in this example require the legislature to permit—or at least not forbid—the extension of such public norms to private firms.

The measures above are familiar and depend mostly on government oversight. Supplementary mechanisms could rely on a more diverse set of participants in the contractual regime, however, so that the entire oversight burden would not fall to government. Examples might include relying on private accreditation to certify compliance with standards, building in third-party participation in the form of an ombudsperson or an intermediary organization that could advocate for consumer interests in the design of service contracts, or requiring private firms to provide opportunities for public participation. In addition, information disclosure requirements could be improved, with a priority placed on producing information in a usable form. Independent auditors might be enlisted to ensure that the information generated is reliable, comprehensible, and widely accessible. Also, contracts could require performance audits by independent third parties. Some supplementary accountability measures might also come from private providers themselves in the form of self-regulation (voluntary adoption of standards and benchmarking) or through management systems that require firms to adopt structural reforms that govern decisionmaking. Lenders, insurers, and investors might be relied upon to place additional demands on firms. These are just a few ideas for alternative accountability mechanisms that could respond to public law concerns.

... [E]xpanding this list will surely require breaking with traditional notions of accountability and mixing and matching measures to produce a combination that works. Rethinking what accountability requires is necessary because public law tends to define accountability as formal and hierarchical, whereas public-private arrangements function as horizontal networks. In the traditional approach, political accountability is envisioned in terms of nested principal-agent relationships: the general public is the principal for elected representatives, elected representatives are principals for a public agency, the agency is the principal for private contractors, and so on. While the public law perspective traditionally has been satisfied with formal accountability of this sort—which relies on the theoretical possibility of tracing responsibility back to a public decisionmaker who can be punished for poor performance—this hierarchical model of accountability becomes more symbolic and less functional with each step away from the electorate. The other primary source of accountability in the traditional view is legal accountability to courts, which suffers from some of the same problems of formality, remoteness, and inaccessibility. Despite their obvious strengths, these sources of accountability also have weaknesses. And as I have argued elsewhere, merely transplanting familiar accountability mechanisms from the public to the private sector, in a one-size-fits-all approach to accountability, might not only fail to protect adequately public law norms, but might also undermine the benefits of relying on private contractors. It might be more productive, then, to think of accountability design as a deeply contextual process, and to imagine it more broadly in terms of measures that spring not exclusively...
from top-down oversight by legislatures, executive branch agencies, and courts, but from a variety of participants—public and private—who are actively engaged on the ground in the contractual regime.

This helps to explain why I define publicization in terms of *extending* public law norms to private actors. I use this word to suggest that private-sector entities could be enlisted in the project of protecting democratic norms, and not just in ways they will resent and resist (though the latter is, of course, possible—there is certainly no dearth of examples of corporate resistance to government regulation). Whether either nonprofit or for-profit firms would enthusiastically embrace these norms is a still more remote possibility, though not unthinkable.

To my mind, the vocabulary of extension evokes a broadly consultative process. It seems more inclusive and multilateral than the language of constraint, and more compatible with the possibility of devising creative approaches to accountability, including instruments that private providers might themselves suggest or generate. Private organizations are familiar with a variety of accountability techniques, even if these techniques have not proven entirely effective: auditing, information disclosure, “total quality management” systems, benchmarking, and the like. Moreover, many industries have long set privately generated standards (even if motivated largely by a concern about government regulation) and have established elaborate standard-setting processes to do so; some even include government-style deliberation. Many standard-setting bodies—the American Correctional Association, the American Medical Association, the National Fire Protection Association, and the American Public Health Association—are longstanding professional associations with considerable credibility. And although market mechanisms, private self-regulation, and independent professional standard-setting have their limits, their mere existence suggests that private firms might have something to offer to the project of instrument design.

The contract negotiation process is promising in this regard. Although contracts between government and service providers take the form of traditional private-law contracts, and need to be attractive enough in their terms to entice private providers to bid for them, these agreements tend to be more unilateral in design than contracts between two private parties: government generally establishes the terms and providers generally agree to them. Thus, there is room to introduce more give-and-take in the process. Indeed, contracts for service provision might benefit from broader consultation with both providers and consumer representatives. As with informal rulemaking, a consultative approach to contract design may inform the agency about a variety of considerations—desired features of the service, possibilities for quality improvements, and opportunities for cost cutting—that it might otherwise have overlooked or designed badly. Such consultation confers additional legitimacy on the resulting contract, too. And although we cannot be certain, compliance with contractual terms might increase if the contracting parties view the process that produced the terms as legitimate. The likelihood of compliance might also increase as a result of greater consultation with contractors themselves, especially if the consultative process is ongoing and explicitly aimed at facilitating performance ex ante rather than simply punishing nonperformance ex post.
C. Skepticism

The notion that privatization can be a means of extending public norms to private actors will invite skepticism. The three principal criticisms I envision are these: the institutions capable of extending these norms to private actors are not motivated to do so; even if they were, publicization will not achieve its aims; and even if it did, the result would be effectively to turn private institutions into public institutions, undermining the benefits of having privatized in the first place. I respond to these objections in turn.

1. Motivation

First, it is hard to predict the level of public demand for extending public law norms to private actors. Most voters might support the idea of publicization in principle if they feared that privatization would result in a loss of due process, public participation, and accountability, though they would not articulate their concerns in these legalistic terms. When privatization runs smoothly, the public appears complacent; when it goes badly, as it recently did during the brief California energy crisis, the public expects government to step back in. As privatization begins to affect politically contentious services (for example, prisons and schools) in a more enduring way, the public may demand greater accountability from, and increased governmental supervision of, private contractors.

If scholarly and judicial attention are any measure of an issue’s political salience, then this prospect seems likely. Private prisons comprise only a small percentage of all prisons and detention centers nationwide, but they have prompted a firestorm of academic commentary and two Supreme Court decisions in a relatively short time. The reaction to vouchers and privately managed schools has been similarly intense, producing a spate of empirical studies, a significant amount of commentary in the popular press, and a recent Supreme Court decision (this one upholding a voucher program). Perhaps the best example of public demand for publicization is the outcry over Health Management Organizations (HMOs). In response to concerns that HMOs deny care in order to cut costs, most states have passed “patients’ bills of rights,” and similar federal legislation has been introduced in Congress. The state laws generally require HMOs to provide some measure of due process before making determinations regarding care or to submit contested decisions to independent review boards.

Still, the skeptics have a point. It is easy to explain why the executive and legislative branches might pursue privatization but harder to predict why they would structure it in the ways I suggest. Although the tools of publicization are available, it is not obvious that the government would be motivated to use them. Indeed, relatively unfettered privatization might be in the interest of both the executive and legislative branches, but for different reasons. For example, a large private contract work force and smaller civil service might enhance executive power. In theory, a president stands to gain by contracting out agency functions because he can appeal to voters as doing more with less and simultaneously diffuse discretionary executive power through a wide range of loosely tracked instruments barely visible to congressional oversight but within the control of political appointees. While fragmenting authority through these mechanisms could make executive oversight somewhat difficult, the president’s political appointees,
with daily control over small-scale decisions relevant to contract management and performance assessment, may be in a better position than Congress to oversee a vast network of contracts, grants, and mandates. This could help the president fend off congressional encroachment on executive power, which most scholars acknowledge has been increasing steadily since the Watergate era.

Members of Congress might favor privatization as a means of bestowing favors on their own supporters and enfranchising favored interest groups to play a watchdog role over agency implementation of legislative deals. Privatization could enable legislators to take credit for program successes in the form of lower costs and improved services—which they can opportunistically draw to the public’s attention—while blaming private contractors, executive agencies, and lower levels of government for program failures, if and when these come to light. In this way, members of Congress can use the contracting process as another vehicle to increase their chances of reelection. This can be done without attracting public scrutiny because members can award benefits to supporters through a web of contracts, subcontracts, grants, and mandates buried deep within the recesses of an agency program. This analysis of congressional motivation relies heavily, of course, on public choice theory, which posits an economic model of the political process in which legislators maximize self-interest. Congress might also delegate broad powers to private actors as a way of insulating discretionary authority from executive control. In their most sweeping formulations, such delegations could raise separation of powers concerns.

It is conceivable, however, that members of Congress might be motivated to obstruct some forms of privatization if doing so will bolster congressional power and limit the executive branch. Congress could of course severely constrain the manner in which executive agencies use contracting through legislative prescription, or more informally through vigorous oversight. A Congress bent on thwarting a particular agency, for example, might encumber agency decisionmaking by requiring a lengthy and detailed review process for grants and contracts, creating layer upon layer of internal review and enabling parties to intervene in agency processes, slowing them to a standstill.

In some cases, moreover, legislators might be moved to oppose privatization out of principle or ideology. Indeed, the most simplified public choice account, which neatly attributes all legislative behavior to an obsession with reelection, is widely considered flawed because it does not account for cases in which members of Congress have legislated to benefit the interests of the diffuse public, even at some personal risk. And even if the simplified account were an accurate depiction of member motivation generally, there are still instances in which an issue’s political salience is so high, and the public so vocal, that a member normally oriented to the concerns of concentrated interest groups will cater instead to the more diffuse interests of her constituents. In some cases, moreover, members are freed of the need to cater to reelection considerations because they occupy relatively safe seats.

So, despite the theoretical plausibility of the claim that members of Congress will support unfettered privatization most of the time, there will be some instances where they will not. Indeed, Congress has already demonstrated a willingness to regulate at least some
kinds of government contracting stringently: the procurement process, by which the
government contracts for goods and services for its own consumption, is subject to
detailed restrictions, including provisions for bid protests and other challenges, which are
adjudicated by the federal Court of Claims. And, as argued earlier, Congress already
conditions federal grants on compliance with a host of regulatory requirements.

Assuming, however, that neither the legislative nor the executive branch will extend
public law norms to private actors, perhaps courts will step into the breach. Again,
skeptics will claim that this possibility seems unlikely. The Supreme Court shows no sign
of, for example, broadly reconsidering state-action doctrine or recognizing implied
private rights of action to allow private parties to enforce federal funding conditions.
However, the Court did find a private doctor under contract with a state prison to be a
state actor for purposes of the constitutional duty to provide adequate medical treatment.
Writing for the majority in West v. Atkins, Justice Blackmun noted that the State bore an
affirmative obligation to provide adequate medical care to West, which it had delegated
to Atkins and which Atkins had “voluntarily assumed . . . by contract.”

There has been some movement in this direction in the lower federal courts. In J.K. ex rel
R.K. v. Dillenberg, a federal district court held that a private regional health facility was a
state actor because it was created exclusively to deliver mental health services to children
entitled to care pursuant to a federal statute. The provider was not merely “doing
business” with the state but executing an entirely delegated responsibility for state health
care duties. Echoing West, the Court held that it would be “patently unreasonable to
presume that Congress would permit a state to disclaim federal responsibilities by
contracting away its obligations to a private entity.” Several other district and appellate
courts have held that constitutional due process applies to private firms when they
assume a state’s mandated duties and are the exclusive service providers. Admittedly,
these are only a few cases, and it would overstate matters to suggest there is a judicial
trend toward cabining privatization. The point is simply that judicial intervention is
possible and that there are signs of it occurring.

While some species of private decisionmaking may not easily submit to judicial review,
as long as there are contracts, regulations, and grant conditions to enforce, courts will be
a possible venue for those seeking to protect public law norms. Indeed, despite the
doctrinal hurdles and obstacles to judicial access that remain, courts are arguably the
most likely of the three branches to extend public norms to private actors because they
are the most likely to be sympathetic to public law concerns. After all, due process,
rationality, equality, public participation, and openness are legal norms that courts
routinely enforce through constitutional adjudication and statutory interpretation. The
judiciary, trained in the rigors of procedural regularity, arguably has greater institutional
competence to address these norms than do the other branches. Whether or not they are
inclined to do so is, of course, another matter.

Moreover, as a function of constitutional structure as well as incentives, courts may be
more disposed to extend public law norms to private actors. The legislature’s
constitutionally assigned role is to tax and spend—to occupy itself with policymaking
and budgeting. The executive must use appropriated funds to implement those policy
decisions effectively. Both of these branches face incentives to focus on productive efficiency, that is, accomplishing social policy goals at the lowest possible cost. The executive branch in particular must determine the best way to provide the social services that Congress funds. And because of electoral discipline, both the legislative and especially the executive branch (which makes the front-line decisions regarding implementation) face incentives to cut costs. Courts, however, do not. The judiciary’s formally assigned role of legal interpretation could not be further from the world of productive efficiency. And insulation from electoral politics lets courts impose public law values on the other branches. Without delving into the long history and rich literature on the “counter-majoritarian difficulty” posed by an unelected judiciary, the claim here is simply that, partly because of their constitutionally assigned roles, courts may be more likely than the legislature or executive to extend public law norms to private actors.

2. Publicization Will Not Work

There is not much to say in response to this objection except to counter that it might. Granted, there are weaknesses in every accountability mechanism considered here, obstacles to their effective use, and anecdotal accounts of severe conflicts of interest. The most carefully designed accountability instruments will have little effect in practice without the political will, funding, and capacity to make them binding. If government is corrupt and inept, even the most carefully designed privatization effort will fail. Moreover, even assuming that public officials have the will to scrutinize private contractors closely to ensure compliance with public law norms, effective oversight might be hampered by limited resources. And in some cases, poor oversight might be the result not of budget shortfalls but of simple inexperience. For example, government agencies that are accustomed to auditing contractors to control fraud may never have developed the capacity to monitor quality. Both of these problems are potentially remediable, but it takes effort.

Undoubtedly, because of such weaknesses, there will be instances in which the paper assurances of cost and quality diverge considerably from actual performance. But while this ought to keep us from blind enthusiasm and complacency, such problems arise with both public and private service provision. The possibility that delegates will shirk their responsibilities and not live up to their commitments is endemic to principal-agent relationships wherever they arise: in corporations, administrative agencies, and legislatures. For privatization to deliver on its promise, we will need a greater political commitment to oversight of both cost and quality than we have today. But while this is a reason for vigilance, it should not, by itself, be a reason for preferring government provision to private provision in all cases.

It is possible, too, that courts will impose limits on the federal spending power, making it more difficult than I have suggested here to extend federal policy goals and public norms to realms that are typically beyond the federal reach. In the wake of the Supreme Court’s decision in *U.S. v. Lopez*, Congress certainly might resort to a more aggressive use of its spending power, and courts may well scrutinize such measures to ensure they do not unconstitutionally coerce states. Even if this occurred, however, presumably Congress
would retain substantial room to maneuver. And surely Congress would remain largely free to impose spending-related conditions directly on private contractors.

Of course, where publicization consists of conditions imposed through voluntary instruments, like loans, grants and vouchers, states and private contractors simply may decline the funds and opt out. This occurred in the health care industry when public hospitals receiving federal funds balked at community service requirements demanding that they provide uncompensated care to Medicare and Medicaid patients “without discrimination.” As a result, many hospitals (both public and nonprofit) restructured as private, for-profit corporations. One can imagine, as well, private religious schools declining to accept vouchers if doing so would obligate them to observe a host of federal requirements. So there will be practical, if not constitutional, limits to the utility of the federal spending power to singlehandedly accomplish the extension of public law norms to private actors. Even acknowledging these limitations, however, the federal purse is a powerful tool.

Perhaps much of the skepticism about publicization is really skepticism about the potential of for-profit firms to serve the public interest. One might argue that the profit motive is simply incompatible with certain policy goals. Under no circumstances, one might claim, will private for-profit prisons actually protect public law norms. How could they, when the profit motive creates incentives to enlarge the prison population and a rational criminal justice policy ought to seek to reduce it? Publicization efforts may be insufficient to overcome such misaligned incentives. And even where the incentives are not so askew, the inevitable difficulties of specifying performance objectives in a contract will always leave slack for providers to exploit, which will always lead to the erosion of public law norms.

Thus far, I have not distinguished between nonprofit organizations and for-profit firms, but it is plausible that the former can be more easily harnessed in the pursuit of public goals than the latter. As Horwitz demonstrates in her empirical study of nonprofit hospitals, nonprofits provide critical public and private goods that conventional markets fail to produce, and the reason for this may be linked specifically to the nonprofit corporate form. The reason for the difference in nonprofit behavior remains unclear: perhaps nonprofit employees assume special ethical obligations; perhaps the law’s historical treatment of nonprofits as serving the public interest has an effect. Perhaps because of a combination of these reasons and others we have yet to identify, nonprofits will embrace democratic norms more readily than for-profit firms. Most importantly, that nonprofit corporate form is positively correlated with different behavioral outputs, as Horwitz shows, suggests that differences in corporate form might matter and that we ought to at least consider them when designing publicization mechanisms.

However, the fact that it might be different or more challenging to extend public norms to for-profit firms does not make it impossible. In the end, one’s optimism or pessimism on this score depends largely on one’s predisposition to the market and government. If one believes, as Harry Arthurs puts it, that “private—especially corporate—power, despite its positive wealth-generating contributions, is also the greatest threat to . . . democratic rights, . . . social relationships and personal dignity,” and if one skeptically views public-
private partnerships and claims of collaboration in light of a long process of corporate resistance to state regulation which dates to the nineteenth century, then one might dismiss any project aimed at enlisting the for-profit private sector in public service delivery more broadly as both naïve and dangerous.

However legitimate the historical account that grounds this view, there is another historical account that is equally powerful: the long tradition of a combination of public and private entities, together providing a variety of publicly desired goods and services that neither could effectively provide alone. As stated earlier, many of the services we now take for granted as state obligations (for example, education, welfare, and even the police) were once privately provided, and a variety of “intermediary” institutions such as nonprofits and religious and charitable organizations emerged early on to provide those services with public-goods dimensions that were difficult for the private sector to provide. The story of the expansion of government services in the twentieth century is in part a story of standardizing, secularizing, and democratizing service provision through expanded regulation of these private-sector players while at the same time multiplying and diversifying those services that the government felt obligated to see provided. This required the development of complex, multilayered relationships between governments at all levels and private contractors (both for-profit and nonprofit), resulting in an interdependent network of actors connected to each other by, among other things, funding, contract, practice, and mission.

Against this background of deep public-private interpenetration, it seems both premature and drastic to suggest that the current trend toward privatization will, as a general and absolute matter, compromise public law norms, regardless of the context in which it is being tried, the availability of accountability mechanisms, and the intentions of government. And it seems unrealistic to think that direct government provision of services is a panacea or even, at this point, a viable alternative. If nothing else, we ought not romanticize the capacity of public agencies, without very significant reform, to directly produce the services and perform the functions that are now provided by this network.

3. Publicization Will Turn Private Firms into Public Agencies and Undermine Gains from Privatization

This final objection is both serious and hard to answer. Adherence to public law norms might be costly for private providers, and those costs might undermine the potential for efficiency gains to some extent. Beermann has a point when he argues that for privatization to achieve its promise of dramatic cost reductions, it would have to be accompanied by deregulation. But it is far less obvious that selectively adding due process, public participation, or oversight will undermine all of the economic gains and technical innovations that might come from reliance on private service providers. Surely it depends on the instruments we use. While publicization will lead to different outcomes than if private providers were entirely unfettered, it remains to be seen just how costly publicization will be. The task now is to describe the circumstances in which some measure of publicization makes sense, and to weigh the tradeoffs that will inevitably arise when choosing among the available publicization tools.
Some of the accountability mechanisms discussed above, such as information disclosure, might be less costly to implement than others. Some measures may effectively pay for themselves. For example, greater attention to contract design through a consultative process, which might be costly on the front end, might produce better ideas about how to provide services effectively and at lower cost. A more careful and inclusive drafting process could conceivably save costs down the line by reducing to some extent future conflict over the meaning of contractual terms. While some contractual incompleteness is unavoidable, many badly drafted service contracts would benefit from greater clarity and specificity. Similarly, enlisting third-party auditors for oversight creates monitoring costs but might be less costly than relying directly on government monitoring. And to the extent that monitoring ferrets out fraud and noncompliance with contractual terms, it can prevent even greater losses over time. This is not to suggest that publicization will be costless, just that it need not entirely compromise the purported benefits of privatization.

Nor will it turn private firms into public agencies. While the extension of public law norms might change how private firms operate, this is nothing new and it is consistent with the profit motive. Private-sector firms are already heavily regulated for a variety of other purposes, such as workplace safety, environmental impact, and employment discrimination. Compliance with such regulations alters firm practices and can be costly but surely does not transform firms into public agencies. Concededly, extreme publicization could produce such an effect—if, for instance, a private firm were required by statute or as a condition of a grant to adopt all the civil-service protections, procedural rules, and sunshine laws governing public agencies, and if courts treated it as a state actor. To become like public agencies, in other words, firms would have to be dramatically and systematically reorganized and reoriented, which seems far-fetched. Surely both nonprofit and for-profit firms would balk at competing for service contracts under such conditions. And I suspect that we can fall well short of this dramatic result and still adequately protect public law norms.

IV. THE STRONGEST CASES FOR PUBLICIZATION

Capitalizing on privatization while reassuring adherents of the public law perspective depends on our ability not only to think inventively about prospective tools of publicization, but also to use them judiciously.

On the one hand, multiple accountability mechanisms would, in Braithwaite’s terms, provide some protection against “gaming” and “fixing” behavior by contractors. So there may be some value to redundancy. On the other hand, these mechanisms are costly and labor-intensive; designing and implementing them effectively would consume considerable financial, political, and intellectual resources. Thus, the question remains: in what instances should we seek to extend public norms to private actors?

First, however, a word of caution: the desire for accountability is harder to satisfy in the context of public-private partnerships, when the government role in service provision is more indirect than when government itself finances and delivers the service. This is because, as Posner explains, authority is typically more diffused in these cases, and the decisionmaking regime includes many players (different layers of government and a host
of private actors), each with their own political power base, goals, and interests. Third-party partners charged with implementation may have significant leverage in policy formulation and implementation, giving rise to the concern that, among other players, big states and powerful providers will enjoy disproportionate influence in the process. Also, as mentioned above, because many government tools, such as loans, grants, and vouchers, are voluntary, governments must be careful to design them to encourage maximum participation.

Another difficulty, Posner points out, arises from lack of coordination across different tools of public action. For example, a substance abuse or child care program may involve conditions imposed by federal, state, and local grants, in addition to vouchers and loans, but the program may lack a concerted strategy for coordinating different goals and requirements. The implementation process for virtually any service or function can be so complicated, moreover, that it becomes hard to simply follow, let alone manage and control. Consider the welfare block grant described earlier, or Salamon’s example of mental health services in Arizona, which are funded by federal and state grants and provided by private contractors—themselves chosen by a private contractor with whom the state, by “master contract,” has contracted out the contracting out.

Public administration scholars such as Salamon and Posner have adopted a “tool selection” strategy to address such problems. They argue that matching the right government tool (for example, loan, voucher, direct regulation) with the right public purpose is essential to a program’s success. Salamon has provided, in this regard, a useful set of dimensions along which to distinguish among different government tools (degree of coercion, directness, automaticity, and visibility) and a set of criteria for evaluating the suitability of different tools for different tasks (effectiveness, efficiency, equity, manageability, and legitimacy). So, for example, the more coercive a tool (such as direct regulation), the higher the price in terms of efficiency. Coercive tools can also be more difficult to manage.

For our purposes, the tools focus is important to keep in mind because it reinforces the intuitive notion that program design can determine program success. It is possible to handicap any scheme badly at the outset by not building in the right incentives and sanctions. But the tools focus assumes that we have already addressed the task I focus on here: identifying those instances in which the extension of public law norms to private actors is most pressing.

This Part describes three considerations that affect the desirability of publicization from the public law perspective: (1) the relative precision with which a service can be specified and the extent of the provider’s discretion, (2) the potential impact on the consumer, and (3) the government’s motivation for privatization. The strongest cases for publicization from the public law perspective involve highly contentious and value-laden services that are hard to specify and over which providers have significant policymaking discretion, that affect vulnerable populations with few exit options and little political clout, and for which the motivation for privatization is discernibly ideological rather than pragmatic.
Each of these considerations has a rough economic analogue: the first overlaps considerably with “contractual incompleteness,” the second might be lumped into what economists count as “quality,” and the third might be expressed as a concern over “patronage and corruption.” While these are not perfect analogues, and the two sides diverge in important ways, there is nonetheless considerable convergence between the economic and public law perspectives on the conditions under which privatization is risky.

Conceivably, then, there might be substantial agreement over cases in which we should never privatize, cases in which we should do so with little hesitation, and hard cases in which some publicization might be preferred to abandoning privatization altogether. If this is true, the question for those hard cases would be: how much publicization, at what cost, and using which mechanisms? Answering this question would require adherents of both the economic and public law perspectives to specify their criteria and explain the trade-offs they are willing to make. By explaining the considerations that matter most from the public law perspective and linking them to similar considerations that come from the economic view, I hope to take the first step in that direction.

A. Ease of Specification and Degree of Discretion

Some tasks are relatively easy to specify in a contract, whereas others defy the most scrupulous attempts at clear definition. For example, specifying what constitutes a quality education is more difficult than defining satisfactory road repair or quality waste collection. This difficulty is unavoidable when a task is highly complex and when reasonable people can easily differ over what constitutes quality, especially with functions or services that implicate deeply held beliefs and involve contestable value judgments. For example, some believe that education should focus on basic skills, while others believe it should prioritize citizenship values. Some parents are fans of standardized tests as a measure of performance, while others believe that only a portfolio of student work can accurately measure progress. And what, precisely, should classroom education encompass? Have students who score well but fail to develop good social skills or high self-esteem received a quality education?

Additionally, specifying a task can be difficult when it involves competing goals that are hard to accomplish simultaneously. For example, we might expect prisons to punish prisoners not only to register society’s disapproval, but also to reduce recidivism. We might expect welfare benefits both to provide a social safety net and to encourage recipients to move from welfare to work. Sometimes these goals are compatible but hard to achieve in equal measures. Specifying the task in such instances requires clear priorities. Although it may be difficult to specify quality for such services, many people would insist that they can assess quality after the fact, once the task is performed or a service delivered. They know quality education, they might say, when they see it.

From the public law perspective, the inability to specify a task because it is value-laden, politically contentious, and complex militates in favor of government provision or very strenuous publicization efforts. Vague contracts leave contractors with considerable flexibility to make judgment calls, trade-offs, and policy decisions to fill in the
contractual gaps. In the public law view, filling these gaps should be entrusted to government, which is bound to make policy decisions in an open and accountable way, in adherence with public law norms. Public law scholars fear that private contractors are more likely to exploit these gaps in favor of cost-cutting measures by, for example, eliminating expensive programs in schools, reducing quality of care in hospitals, or hiring unqualified guards in private prisons. They are also concerned about the political legitimacy of conferring policymaking discretion on nongovernmental actors.

From the public law perspective, unconstrained discretion is dangerous because it can lead to abuses of power. Where discretion is broad, therefore, and contractors make policy decisions rather than merely implement the choices of politically accountable bodies, the case for publicization, all things being equal, is strengthened.

Of course, public law scholars have long conceded that policy decisions inhere in virtually all instances of implementation. When Congress delegates authority to an executive agency, for example, it always gives the agency some discretion to interpret the terms of the delegation and the substance of the statutory mandate. The implementation process translates abstract legislative choices into meaningful burdens and benefits. For instance, whether someone qualifies for a statutory entitlement can depend on an eligibility test determined through a long administrative process involving rulemaking, interpretive rules, guidance documents, opinion letters, and ultimately, individualized adjudications. Implementing Congress’s policy choices involves multiple discretionary moments during which one or more decisionmakers refine vague terms, transforming them into actual determinations that affect people on the ground.

But the public law view nonetheless draws a strong distinction between the discretion to make standards and the discretion merely to implement them. Determining eligibility standards for a benefit like Medicare involves far greater discretion than merely implementing eligibility standards determined exogenously. By the same token, the discretion to define what constitutes a quality education is greater than the authority to implement detailed directives defining curricular content. The degree to which public law scholars are concerned about discretion depends in part on the complexity, contentiousness, and value-laden nature of the service, as well as on its potential impact on consumers (a consideration I explore in more detail below). For example, determining how precisely to collect solid waste leaves open a narrower range of options with fewer salient consequences than does the day-to-day discretion that correctional officers exercise over whether inmates receive medical care or are assigned to administrative segregation for disciplinary infractions. In part, this turns on the intuition that less is at stake in decisionmaking concerning waste collection than in decisionmaking concerning incarceration.

While all instances of privatization involve some discretion, and one cannot distinguish scientifically among degrees of discretion, one can still plausibly differentiate broad policy discretion (for example, authority to design a program entirely “in the public interest”) and inconsequential policy discretion (for example, authority to collect garbage on any three nonconsecutive days of the week). From the public law perspective, the
broaden the discretion afforded the private decisionmaker, the greater the need for accountability, rationality, and procedural fairness.

Economists would call the difficulty of specifying quality “non-contractibility” and express similar concerns in different terms. For example, Hart, Shleifer, and Vishny refer to gap-filling as the allocation of “residual control rights.” When quality is hard to specify, it becomes harder to prove that the provider has fallen short of contract specifications. Moreover, non-contractibility can create incentives for providers to cut costs at the expense of quality because the quality loss may not technically violate the terms of the contract. This is known as the “quality-shading hypothesis.” Where quality is non-contractible, government will likely measure performance in terms of inputs, such as time and materials invested in a task, rather than outputs such as product quality, which means payment will be “based in large part on activity rather than on results.”

But the economic perspective is not as concerned as the public law view with the legitimacy of conferring policymaking authority on nongovernmental actors. Indeed, non-contractibility is portrayed as problematic primarily because it could compromise service quality and create costly monitoring problems, not because of a concern about democratic accountability. Nonetheless, both perspectives agree that privatization is riskier when quality is harder to specify. Some would counsel against privatization where non-contractibility is severe. At a minimum, however, the case for publicization is stronger for hard-to-specify tasks that confer considerable policymaking discretion on private contractors. In such instances, we might wish to ensure that contractors fill gaps in an open and accountable way.

B. The Potential Impact of the Service on the Consumer

Some services involve policy decisions and raise normative questions about which people care deeply, whereas other services are more routine in the sense that they tend not to raise questions of individual rights or fundamental fairness or substantially affect the more important conditions of life.

This standard is of course imprecise: one person’s routine service might be crucially important to another. Nevertheless, it seems plausible that services that directly and substantially affect health, liberty, safety, and personal autonomy should qualify as important and nonroutine.

For example, highway maintenance, solid waste collection, and mail delivery seem routine. While these might be difficult tasks which involve important judgments and affect human health and commerce, they are not especially value-laden activities and do not generally implicate our most cherished freedoms and aspirations. Frankly, they are hard to get excited about. While some might prefer government to provide these services, and while it might matter symbolically whether the people in uniform fixing the street, collecting waste, or delivering mail actually work for the government (my affection for the Fed Ex delivery guy is slightly less than it is for Fred the mailman, though I have friends who feel differently), for most people, most of the time, cost and technical efficiency matter most.
From a public law perspective, these comparatively routine services are not strong candidates for our most strenuous publicization efforts. This is not only because they do not tend to implicate fundamental rights or affect the most important conditions of life, but also because they can be relatively easily specified and monitored so they pose fewer non-contractability problems. Nonetheless, even for routine services, the public law concerns about fairness and nonarbitrariness suggest the need for some minimal extension of public law norms to private contractors. For services like waste collection and mail delivery, as with utilities like water, gas, and electricity, we might demand that contractors provide universal access, comply with antidiscrimination norms, and put procedures in place to prevent arbitrariness in termination decisions. This is especially important when the provider has a monopoly, a point with which economists, for whom the absence of competition augurs ill for the success of privatization, would likely concur.

There are other services that are decidedly nonroutine and that affect consumers in more fundamental ways. These tend to overlap significantly with those services for which quality is difficult to specify. The public law concern about observing public law norms is especially great, however, where the populations most directly affected by privatization are vulnerable. This vulnerability might arise because these populations lack resources and political clout, enjoy few viable exit options, or are relatively invisible to the population at large—such as prisoners, detainees, persons involuntarily committed to institutions, welfare and disability-aid recipients, Medicaid patients, and school-age children.

Vulnerability can be exacerbated when the consumers of the service are not the same as the payers—when taxpayers, for instance, finance prisons occupied by convicts, welfare received by eligible low-income applicants, and schools occupied by other people’s children. In such instances, privatization may present a choice between taxpayers’ and consumers’ interests. For example, taxpayers who finance the services may prefer to conserve public monies while consumers might be more interested in improving the quality of the services. Taxpayers might be disinclined to invest in better prisons, more generous welfare, and more costly drug treatment, HIV prevention, and job-training programs. They might be willing to compromise prisoner safety and medical care to conserve resources. Prisoners will likely think differently.

This conflict disappears, however, when those who pay for the service also consume it. The potential for such conflicts is greatest with despised and dependent populations whom payers may regard with little sympathy. So again, the vulnerability consideration has an augmenting effect here. Where such conflicts between constituencies arise, the case for publicization, from the public law perspective, will tend to be stronger—in large part because payers may wield more power than consumers, who may suffer grave injury or loss as a result.

From the economic perspective, the concern about privatization’s impact on the consumer—specifically the concern about vulnerable populations that might suffer great harm—might count as part of the “quality” of the service. But economic accounts of privatization rarely define quality specifically to include public law concerns, and
certainly not in the normative terms used here, so it is difficult to tell. Conceivably, however, adherents of the economic perspective might be willing to accord weight to such considerations. Undoubtedly, they would be sympathetic to the monitoring costs that arise when the interests of consumers and payers conflict. And they would agree, as argued above, that the absence of competition can be problematic, not primarily because it might lead to abuse of power by private actors with monopoly power, but rather because it undermines the potential for cost savings. The public law emphasis on the risks vulnerable parties face may be a point of divergence between the two perspectives: when the consequences of privatization are disproportionately borne by vulnerable populations with few or no exit options, the public law view argues especially vigorously in favor of maintaining due process and judicial review, which traditionally apply only to government actors.

C. Motivation

From a public law perspective, the motivation behind the government’s decision to privatize would likely influence judgments about the extent of publicization required. Because they are attuned to the implications of privatization for democratic norms, adherents of the public law view would be especially suspicious if government privatizes primarily to avoid its constitutional and legal obligations. Discovering that this, and not the pragmatic desire to cut costs and improve quality, is the impetus behind privatization would militate in favor of publicization.

It may be impossible to detect government’s true motivation for privatization, and in many cases that motivation may be mixed. However, if an ideology of avoidance does emerge as the primary motivation, the public law view counsels caution, especially when a single instance of privatization is linked to a larger program of privatization that might intentionally and cumulatively seek to erode public law protections. Privatization on a broad enough scale could lead to a compromised, even anemic, government rather than to the extension and proliferation of democratic norms that adherents of the public law view support. From this perspective, successful privatization requires a strong government capable of overseeing and, ideally, partnering successfully with the private sector to deliver high-quality services at a reasonable cost. If, instead, widespread privatization becomes a justification for shrinking government, there could be a tipping point beyond which government oversight and enforcement capacity will be so badly eroded as to disable it from playing a meaningful role in governance.

The economic perspective may share the concern that privatization be economically justified, but would likely frame this concern as a worry about corruption and patronage rather than ideology. As Hart, Shleifer, and Vishny explain, the presence of corruption in government will create a bias toward privatization, whereas the presence of patronage will create a bias toward public provision. The case for privatization is weaker, therefore, when government is corrupt. Moreover, the economic perspective demands adequate enforcement of contractual terms, so ideological considerations that result in less monitoring of contractors’ performance, or grants of monopoly power, may also weaken the case for privatization. Thus, both perspectives are sensitive to the need for privatization to be justified by pragmatic concerns about cost and quality. It seems
plausible that adherents of both views would support publicization when these pragmatic motivations for privatization are weak or absent.

From the public law view, these three considerations—the ease of specifying goals and the extent of discretion, the impact on the consumer, and the motivation for privatization—strongly influence the case for extending public law norms to private actors. When choosing where to invest our energy, we should tend not to publicize services that are relatively easily specified and monitored via contract, that involve only limited discretion, that are not especially value-laden, that tend not to involve vulnerable populations, and that are motivated by pragmatic concerns. The best illustration may be municipal infrastructure maintenance—waste collection, transportation, and the like.

Administering a welfare program, operating a mental health facility, running a school, or managing a prison, however, is different. These are intensely value-laden activities. They involve direct contact with, and sometimes custody of, vulnerable people. They require discretionary decisionmaking that could have a potentially enormous effect on individual health and welfare. Moreover, these nonroutine tasks are hard to define, and conflicts between the paying and consuming constituencies can lead to pressure for cost cutting at the expense of legal protections.

The most difficult cases will arise when these considerations pull in different directions. On the one hand, water and electricity provision are services on which both highly vulnerable and less vulnerable people are heavily dependent; the deprivation of these services would have a significant effect on consumers and would interfere with their capacity to perform daily activities. On the other hand, the delivery of water and electricity does not involve as much discretion as other services. What publicization efforts should attend these services? From the public law perspective, are utilities more like education and incarceration than like waste collection? And, if forced to choose, which of the considerations discussed above is most important: the absence of exit options, the involuntariness of the consumer’s situation, or the degree of the provider’s discretion? To answer difficult questions such as these requires agreeing on the relative weights that each of these considerations should have and developing a framework for making trade-offs among them. This would be a productive next step for economists and public law scholars to take together.

The economic perspective is concerned with many, but not all, of the things that matter to public law scholars. There are public law analogues for many of the economic reasons that militate against privatization: for example, when principal-agent problems (including adverse selection and moral hazard) are severe and monitoring costs high, when actual or potential competition is low, and when there is corruption.

The two perspectives overlap to some extent as well in their choice of remedy. Because contract can help minimize some of these problems, the economic perspective counsels careful drafting. Well-drafted contracts specify tasks in as much detail as reasonably feasible, establish performance goals, and provide for sanctions in the event of nonperformance. Though perfect specificity might be both too costly and practically impossible, careful articulation of contractual terms would help to reduce uncertainty and
provide a basis for ex post monitoring. Contract is important to the public law scholar as an instrument for reducing the slack in the principal-agent relationship, but it plays an additional role, potentially enfranchising consumers to participate in the contractual process and hold private providers to account for their decisions. This is something public law scholars value because of its implications for democratic accountability. There is no perfect analogue with which to express this view in economic terms. Nonetheless, there might be considerable agreement between the economic and public law views about the importance of clear and enforceable contractual terms to the success of privatization.

Clearly, however, the convergence is imperfect. Each perspective has its menu of concerns, accessible in a different language and stemming from different cultural and normative frameworks. Some of these concerns are incommensurable, and there will be resistance to giving them up. And yet there might be substantial agreement about those instances in which publicization is warranted, and perhaps even some overlap in which tools to use when. This further task of elaborating that commonality is an endeavor from which both perspectives might profit.

V. Conclusion

In this Commentary, I have suggested that privatization, far from weakening democratic norms of due process, rationality, equality, and accountability, could instead extend these norms to private actors—an effect I call “publicization.” The instruments for accomplishing this extension are plentiful, I argued, and already, to a significant extent, in use. They include direct regulation, conditioned funding, contract, and tort liability, among other things. In making the case that publicization is possible, I have sought to assuage the concerns of those who object to privatization out of fear that greater reliance on private service provision will necessarily erode public norms. Indeed, I proposed that the reach of government might expand, through privatization, into areas and matters that are constitutionally or practically otherwise off-limits to government. In addition, to bridge the gap between those who, from the “public law” perspective, oppose privatization and economists who tend, out of a preoccupation with cost and quality, to support it, I explained that there is in fact considerable convergence between the two groups about when privatization is most risky, and when, therefore, it makes the most sense to build publicization measures into privatization schemes.

This Commentary is, of course, only a first step. Even if we can agree on those instances that call for extending public law norms to private actors, precisely how to do it in particular cases without stifling the creative and productive capacity of private actors remains a significant challenge. I have suggested here a variety of mechanisms—some more coercive than others, some emanating from public agencies and others from private parties—for doing so, and I have portrayed this endeavor as an exercise in which private actors, both nonprofit and for-profit, must be themselves enlisted. I have tried, above all, however, to bridge a gap between the economic and public law view that I believe to be smaller than it seems in order to narrow areas of disagreement and focus the debate on the hardest questions. Recognizing that there is much left to do, if I have succeeded even modestly in this regard, I will consider the effort to have been worthwhile.
§ 6.4. Case Studies


I. The Role for Government in Airline Security in the United States

Before 9/11, private airlines were assigned the task to screen passengers and baggage at airports as part of an overall service responsibility which included ticket agents, pilots, flight attendants, and maintenance personnel. They operated under FAA oversight. The airlines did not actually employ the screeners as they did flight and maintenance staff; instead they entered into contracts with private security firms. Not surprisingly, these contracts went to the lowest bidders. As a result, the private security firm employees were likely to be poorly trained, lacking in language skills, and otherwise unqualified (they were sometimes convicted felons). The Department of Transportation Inspector General found in the years prior to 2001 that undercover agents could penetrate security at most airports. Indeed, just days after 9/11, seven of 20 people carrying knives passed through security at Dulles Airport.

It was against this background that Congress confronted the need to improve airport security after 9/11. Senators McCain and Hollings introduced Senate Bill 1447 to federalize the airport security workforce. The bill required new training and performance standards for the public employee screeners and also reconfigured the stationing of law enforcement personnel at airport checkpoints. After initially supporting the Senate bill, President Bush shifted his support to an alternative House bill designed to limit the federal role. This bill would have installed federal supervisors at baggage and passenger screening checkpoints, but left the security workforce itself—the inspectors of passengers and baggage—in private hands. Ultimately, support for complete federalization in Congress was so overwhelming that the President backed off the partial privatization bill.

The Aviation and Transportation Security Act (ATSA) was passed by Congress and signed by the President on November 19, 2001. It created a new department, the TSA, within the Department of Transportation, and gave the TSA responsibility for aviation security. The TSA was to federalize airport security screeners within one year. All screeners were to speak and read English, be U.S. citizens, have no criminal records, and have graduated from high school. Screeners were also required to complete 40 hours of classroom instruction and 60 hours of on-the-job instruction. The act also provided for a pilot program to test private security personnel in up to five airports across the country. In addition, airports were to be allowed to apply to reprivatize (as part of the Security Screening Opt-Out Program) after November 19, 2004. . . .

22 The public employment requirement permitted five airports to opt-out: San Francisco, Kansas City, Rochester, Jackson Hole, and Tupelo.
Ultimately . . ., Congress, in passing a public employment bill, succumbed to Democratic arguments that U.S. airports, like borders, should be patrolled by federal personnel. Over 60,000 new federal employees were created. Despite existing counterexamples, the notion of private contractors conducting safety inspections struck many as a distortion of public responsibilities. In comparing airport security officials to custom officials, Congress in effect endorsed the requirements of public service. Federal officials are different from private contractors—they take oaths and wear badges, although this requirement did not hold for TSA inspectors who do not in fact have the power to arrest. Even though the cost of airport security officials was paid by air travelers, not the federal treasury, their public status was equated with the proper exercise of public authority. This response countered the trend in the United States towards privatization of government services, and it countered trends in other countries with greater commitments to public sector solutions.

The use of public officials to provide security at airports reflects unease about the competence and even the loyalty of those employed to perform the inspection task. The requirements of no criminal records and higher educational and training levels are directed at assuring more reliable employees. The requirement for U.S. citizenship presumably filled the loyalty gap, although one might wonder why that requirement should not also apply to members of the United States military. These requirements might have been stated contractually if the airport security personnel were employed by private contractors. Since the privatization trend has already encompassed prison guards, military officials, and others who are equally part of the security network, it does not seem compelling to curtail it in this one instance.

So something more is at work here. President Bush, in his signing message, noted that “[f]or the first time, airport security will become a direct federal responsibility, overseen by [a] new undersecretary of transportation for security.” This commitment to public solutions was part of the President’s need to assure the public that airlines were secure. A public solution had symbolic value—it provided assurance to a skeptical and even fearful traveling public that the government was in charge.

Of course, public status itself is not sufficient to assure superior performance. And, as a conceptual matter, carefully supervised private employees would seem to be capable of effective inspections. . . . [P]ublic service dictated the outcome for political reasons without a careful analysis of the nature of the governmental functions involved. Such an analysis would have explored the jobs themselves to determine whether they were “inherently governmental” and not subject to “competitive sourcing.” These terms are employed under the government’s A-76 competitive sourcing process, where agencies like Homeland Security are required to designate what jobs are eligible for contracting out.

If such a study had been undertaken by Homeland Security, the airport security function would have been studied more carefully. Dividing jobs into levels and components would have helped identify roles that should be public and private. This process would have likely rendered those inspectors closest to the traveling public as most eligible for private contracting. These are the TSA officials whom we all encounter in our air travels, and
their functions are limited in ways that make them readily contractible to private hands. In so doing, however, the goal of reassurance of the public might not have been achieved. At the time, public status—the presence of an inspector with government authority—was deemed necessary to achieve the overall goal of the program. Publicization, therefore, reflects values that transcend the actual nature of government jobs. In effect, when privatization occurs, symbols of authority and accountability that public status exemplifies cannot be delegated. This does not mean that private contractors do not have a role to play in clearly defined assignments (such as airport inspections), but it does mean that some duties by those supervising them may be inherent in sovereignty. It is entirely possible that the United States erred on the side of excessive publicization in the airport security setting for reasons that go to public assurance rather than public control. That at least is what the European experience suggests.

II. PRIVATIZATION AND AIRPORT SCREENERS IN EUROPE

During the debate over the airport security bill in the United States, House members arguing in favor of private inspectors noted that European countries used private inspectors at airports. Since it is assumed that European countries are more comfortable with public employment than the United States, this argument came as a surprise. Statist solutions may be more congenial in Europe, but privatization has a significant role to play as well.

Airport security is largely a member state decision in the European Community (EU); the Community’s role is limited to setting common standards for airport security. . . . The Community standards cover use of airport security equipment, the requirement of background checks for security personnel, and provision for unannounced inspections of individual airport security arrangements. The EU, however, does not dictate member state choice between public or private providers of security.

In some states the function of baggage and passenger inspections has been delegated to private actors. German law embraces a concept of “functional privatization” which allows entities to receive assistance from private parties. The delegation to private hands, however, is accompanied by several important limitations: (a) the delegation must be supervised by a public official; (b) administrative procedures necessary to protect individuals must remain in place; and (c) decisions that affect individuals must be made by public officials.

These procedural protections are meant to assure public oversight of the airport inspection process. They are deemed sufficient alternatives to direct employment of screeners at many European airports. By setting common control standards, Community law helps determine when public officials should be required to perform “public” functions.

The use of private screeners in European countries reflects the growth of what Professor Alfred Aman has called the “non state public sector.” He points to the danger of a “democracy deficit” if accountability to the public is ignored. The Community regulations seem to anticipate that concern. They provide for audits and unannounced
inspections of member state airports and contemplate oversight by state employed security control personnel. . . .

III. RECONCILING THE USE OF PUBLIC AND PRIVATE OFFICIALS IN AIRPORT SECURITY

In the United States, . . . [o]fficials performing “inherent government functions” are explicitly excluded from the privatization calculation. The distinction between competitive and inherent functions drawn by the government’s A-76 process, discussed above, is not easy to draw. . . .

. . . [S]uppose the TSA were to put public inspectors up for competitive sourcing under A-76. What would be the outcome? TSA employees would argue that they performed inherent government functions. Are they right?

Security is a role government is designed to perform, but it can still delegate some of that function to private hands. Private prisons are a classic example. It is the level and degree of the security function that tests the proposition. The physical act of searching passengers or baggage is a limited one—it does not involve the power to arrest that would command a government presence. Congress could have voted the other way on TSA officials, at least as to the inspection function performed by those officials. So it does not seem to be the inspection function itself that requires government personnel.

However, placing public oversight and control in the hands of private contractors rather than the Undersecretary of Homeland Security would be a different matter. These functions are likely to be inherently governmental because they involve judgment and discretion. Thus they may be nondelegable under A-76 and under statutory and even constitutional principles.

The requirements of supervision and control are indicators of inherent functions. Ultimately it is the duty of government to govern, and this function may not be transferred to private hands under our Constitution. While the actual inspections are relatively limited and constitute assignments that may be delegated, the responsibility to ensure proper performance by those officials is a different and inherently governmental one. European countries that permit private airport inspectors who are controlled, trained, and overseen by public officials seem to have reached the same conclusion. Indeed, we have made the same choice in the opt-out context for the five airports that were excepted from the Act, since as private contractors they are still controlled by the Undersecretary of Homeland Security.

A. Establishing Public Principles for Airport Security

Where should the lines be drawn? If one were doing a simple make or buy choice, the question would be limited to which program (public or private) performed better and at what cost. So long as performance could be measured, in other words, the case for privatization remains strong. But this is because public oversight is possible. Recent studies do not confer a performance edge on government versus private inspectors. Therefore private contractors have a case to make. But this would not extend to the oversight function itself. Determining to what extent privatized employees are adequately

421

Draft — do not distribute!
supervised and trained is a function that is not easily converted into performance standards. In the United States and Europe these jobs are in public hands precisely because they are not repetitive or reproducible, but are subjective and judgmental. Duties that should be performed by public officials in both settings are those that are difficult to describe or enforce in contractual terms, since they involve matters of judgment, control, and responsibility.

The European Community dealt with these issues in its airport security directives. Congress in setting up the Transportation Security Administration gave it the capacity to achieve the level of oversight needed should decisions be made further to privatize the airport inspection function. In both jurisdictions, however, public controls remain in place to limit the effects of privatization and to monitor its success or failure.

B. Applying Theories of Public Control

Let us assume that airport screeners perform delegable duties, whether to public or private hands. How far up the chain of command can the public role be delegated without abandoning the government’s responsibility to govern? Professor John Donahue has provided some guideposts to help answer this question in efficiency and public management terms. Donahue argues for government to choose the best structure (public or private) necessary to assure accountability. He grants the accountability criterion priority over efficiency in public decisionmaking. If, as has long been recognized, accountability is the necessary condition for effective privatization, then delegations to private hands that ignore its implications are inherently suspect.

When it comes to the provision of airport security, accountability is a crucial variable. Since security is an essential role of government, privatization must be integrated into that framework, not the other way around. This requires a careful analysis of the various functions performed by the officials involved. At European airports, the physical inspections of passengers and baggage are largely outsourced to private providers; in the United States this is generally not the case.

The crucial accountability question is not whether these specific arrangements are outsourced, but whether those performing the inspection functions are properly supervised. Oversight and accountability are the decisive variables. As long as the delegated assignments are clearly limited by contract, and the oversight function is publicly performed, the private interest can be accommodated to the public. In Europe the Community rules make the public oversight role clear. In the United States, the ATSA is also reassuring on this note. Even if it is assumed that the exceptions to public inspectors become the rule in the future, if protections through contractual limitations and public oversight (including clearly defined administrator responsibility at Homeland Security) are achieved, accountability can still be assured.

Unlike many other situations in government today, most prominently the private military in Iraq, airport security has been properly structured as a joint private and public function. In this setting, privatization can have a productive role to play while the national security interest is protected through accountability mechanisms that reflect the inherent role of
government. This leaves one to wonder why it is that we have granted airport security such a careful degree of public control and have left many other security concerns out of the ambit of reform. Airport security achieved a crucial status due to the events of 9/11, but its lessons apply to a broader range of national security matters where privatization often crosses policy lines. . . .


INTRODUCTION

The Federal National Mortgage Association (Fannie Mae) and the Federal Home Loan Mortgage Corporation (Freddie Mac) are the two dominant entities in the secondary residential mortgage markets of the United States. They are an important and prominent part of a larger mosaic of extensive efforts by governments at all levels to encourage the production and consumption of housing.

Fannie Mae and Freddie Mac are a unique part of this effort. Though they appear to be “normal” corporations, each with shares that trade on the New York Stock Exchange, they in fact have federal government origins and entanglements that make them quite special. Indeed, they are often described as “government-sponsored enterprises” (GSEs). Yet, their specialness is a two-edged sword: On one side, they cause interest rates on many residential mortgages to be lower than would otherwise be the case; on the other, their size and mode of operation have created a significant contingent liability for the federal government and, ultimately, taxpayers. In addition, the size and prominence of the two GSEs has recently led to concerns about the larger consequences for the U.S. economy if either were to experience financial difficulties.

There is strong evidence that home ownership has positive spillover (“externality”) effects for society and thus that targeted policies to encourage home ownership (by those who would otherwise rent) can improve a society’s allocation of economic resources. However, the broad policies that encourage home ownership do not address those spillover effects in a focused way (and policies that encourage more rental housing, of course, are contrary to the goal of encouraging home ownership). Instead, they simply encourage the consumption of more housing—at the expense of other things—by those who would have bought anyway, with the consequence that our society’s resources are less efficiently allocated. The encouragement that is provided through Fannie Mae and Freddie Mac is largely of this broad-based nature and thus suffers from this same distortionary consequence.

The special governmental links that apply to Fannie Mae and Freddie Mac yield little that is socially beneficial, while creating potential social costs. Consequently, the appropriate
“first-best” policy would be to privatize them completely—that is, to sever all governmental links and convert them to truly “normal” corporations—as well as to pursue other measures that would better address the positive externality of home ownership and efficiently reduce the cost of housing. In the event that this true privatization does not occur, suitable “second-best” policies are discussed as well.

SOME BACKGROUND

What They Do

Fannie Mae and Freddie Mac each operate two related lines of business: They issue and guarantee mortgage-backed securities, and they invest in mortgage assets. Both businesses warrant further explanation.

Mortgage-Backed Securities. A typical transaction in today’s mortgage markets involves a swap of a pool (bundle) of residential mortgages that have been originated by a commercial bank, a savings and loan (S&L) association, or a mortgage bank for a set of mortgage-backed securities (MBS) that have been issued by Fannie Mae or Freddie Mac and that represent a claim on the interest and principal payments on the same mortgage pool. The two companies guarantee timely payment of principal and interest to the MBS holders and, for that guarantee, charge about 20 basis points (0.20 percentage points) annually on the outstanding principal amounts. The originators, in turn, have a liquid security that they can hold on their balance sheets (with a substantial regulatory advantage for commercial banks and S&Ls over holding the underlying mortgages themselves) or sell in secondary markets (which mortgage banks immediately do). . . . As of year-end 2003 the two GSEs together had more than $2 trillion in outstanding MBS.

Mortgage-Related Assets. Instead of swapping MBS for mortgages, Fannie Mae and Freddie Mac may buy the mortgages outright and hold them in their portfolios (or sometimes securitize them and sell the MBS to the public). The two companies also repurchase their MBS through transactions in the secondary market, and most of their mortgage-related assets are now repurchased MBS. . . . The two companies’ mortgage-related assets at year-end 2003 totaled almost $1.8 trillion. The two companies fund their mortgage-related asset holdings overwhelmingly through the issuance of debt.

Some History

Fannie Mae was created in 1938, under the authority of the National Housing Act of 1934. Until 1968, it was a unit within the federal government. Its function was to expand the availability of residential mortgage finance by buying mortgages from originators and holding the mortgages. These purchases were funded through debt issuances that were direct obligations of the federal government.

As part of the Housing and Urban Development Act of 1968, Fannie Mae was spun off from the federal government and became a publicly traded corporation, but it retained an array of special government features (discussed below). Its function continued to be that of expanding the availability of residential mortgage finance through mortgage
purchases, largely from mortgage banks, that were funded overwhelmingly by debt. Also, Fannie Mae was replaced within the federal government in 1968 by the Government National Mortgage Association (Ginnie Mae), an entity within the Department of Housing and Urban Development (HUD) that guarantees MBS that represent claims on pools of mortgages that are insured by the Federal Housing Authority or the Veterans Administration.

Freddie Mac was created in 1970 also to expand the availability of residential mortgage finance, primarily through the securitization of mortgages purchased from S&Ls. Though the first MBS were issued by Ginnie Mae in 1970, Freddie Mac was a fast second with its initial MBS issuance in 1971. Through the 1970s and 1980s, Freddie Mac was owned solely by the twelve banks of the Federal Home Loan Bank system and by the S&Ls that were members of the FHLB system. Freddie Mac became a publicly traded company in 1989, but with the same ties to the federal government that Fannie Mae has.

Through the 1970s and 1980s the business strategies of the two GSEs were somewhat divergent . . . . Fannie Mae tended to focus on mortgage purchases for its own portfolio (it issued its first MBS only in 1981), while Freddie Mac tended to focus on MBS issuances. Since 1990, however, the two companies’ business strategies have been largely similar: rapid growth of both their portfolio businesses and their MBS businesses. Indeed, their growth rates since 1990—especially for Freddie Mac—have been breathtaking . . . . [T]heir growth rates have been far faster than that of the overall mortgage markets. As of 1980, the two companies’ mortgage holdings plus MBS accounted for only 7 percent of the total of all residential mortgages. By 2003 their aggregate involvement in the mortgage market came to 47 percent . . . .

THE SPECIAL STATUS OF FANNIE MAE AND FREDDIE MAC, AND THE CONSEQUENCES

Fannie Mae and Freddie Mac are not ordinary corporations. They differ from all other corporations in the United States in many ways . . . .

Advantages

- They were created by Congress and thus hold special federal charters (unlike virtually all other corporations, which hold charters granted by a state, often Delaware);
- The president can appoint 5 of the 18 board members of each company;
- Each company has a potential line of credit with the U.S. Treasury for up to $2.25 billion;
- Both companies are exempt from state and local income taxes;
- They can use the Federal Reserve as their fiscal agent;
- Their debt is eligible for use as collateral for public deposits, for purchase by the Federal Reserve in open-market operations, and for unlimited investment by commercial banks and S&Ls;
- Their securities are exempt from the Securities and Exchange Commission’s registration and reporting requirements and fees;
• Their securities are explicitly government securities under the Securities Exchange Act of 1934; and
• Their securities are exempt from the provisions of many state investor protection laws.

These benefits directly lower GSEs’ costs and have also created a “halo” of implied federal government protection for the two enterprises. That halo effect has been reinforced by past government forbearance when Fannie Mae was insolvent on a market-value basis in the late 1970s and early 1980s and by a taxpayer bailout of the Farm Credit System (which had similar benefits) in the late 1980s. Perhaps most importantly, because the financial markets believe that the special GSE status of Fannie Mae and Freddie Mac implies that the federal government would come to their (and their creditors’) rescue in the event of financial difficulties—despite specific language on every security that they issue that declares that the securities are not guaranteed by or otherwise an obligation of the federal government—their debt is treated favorably by the financial markets: They can borrow on more favorable terms (i.e., at lower interest rates) than their credit ratings as stand-alone enterprises would otherwise justify. Typically, they can borrow at rates that are more favorable than those of an AAA-rated corporation (though not quite as favorably as the rates on the debt of the U.S. government itself), even though their stand-alone ratings would be about AA– or less; this translates into about a 35–40 basis point advantage. Similarly, they enjoy about a 30 basis point advantage in issuing their MBS as a consequence of their special GSE status.

**Disadvantages**

• Their special charters restrict them to residential mortgage finance.
• They are specifically forbidden to engage in mortgage origination.
• They are subject to a maximum size of mortgage (linked to an annual index of housing prices) that they can finance; for 2004 that limit for a single-family home is $333,700.
• The mortgages that they finance must have at least a 20 percent down payment (i.e., a maximum loan-to-value ratio of 80 percent) or a credit enhancement (such as mortgage insurance).
• They are subject to safety-and-soundness regulation—for example, minimum capital requirements and annual examinations—by the Office of Federal Housing Enterprise Oversight.
• They are subject to “mission oversight” by HUD, which approves specific housing finance programs and sets social housing targets for the two companies.

**The Effects on Residential Mortgages**

The presence of Fannie Mae and Freddie Mac in the secondary mortgage market influences rates in the primary mortgage market. Their activities cause the rates on the “conforming” mortgages that they can buy to be about 20–25 basis points lower than the rates on “jumbo” mortgages. In addition, their presence may well bring greater stability to the mortgage markets, and historically they were able to bring greater uniformity and unification to what otherwise would have been localized and disconnected markets, since
regulatory restrictions on interstate banking and even intrastate bank branching in some states persisted for most of the 20th century and prevented banks and S&Ls from bringing this unification. Also, the two companies may have been focal points for marketwide standard setting with respect to the technological advances in the processes of mortgage origination. And, historically, they were important in the development of MBS and of mortgage securitization generally as an alternative efficient mechanism for residential mortgage finance.

**THE POLICY ISSUES**

Fannie Mae and Freddie Mac do not, of course, exist in a vacuum. There are at least six larger issues that surround them and deserve greater exploration, so as to evaluate the special position and role of the two companies. Those larger issues are: (1) the widespread public policies in the United States that encourage the construction and consumption of housing; (2) the safety-and-soundness regulation of financial institutions where there are concerns about the social consequences of the insolvency of those institutions; (3) the possible systemic consequences of their size and behavior; (4) the question of who should bear the interest-rate risks concomitant with the long-term debt instrument that is the modern mortgage in the United States; (5) the question of the efficient transmission to homebuyers of the benefits bestowed on Fannie Mae and Freddie Mac as a consequence of their special GSE status; and (6) the question of possible inherent efficiencies or inefficiencies of the two companies’ activities. . . .

**Housing**

U.S. public policy, at all levels of government, embraces extensive policies to encourage the construction and consumption of housing. . . . There is a reasonable theoretical basis for the existence of positive externalities that would support government policies to encourage home ownership. . . .

The logical linkage to policy from this externality would be to have tightly focused programs that would encourage low- and moderate-income households, who may be on the margin between renting and owning, to become first-time homebuyers. Such programs could provide explicit subsidies for reducing down payments and reducing monthly payments.

[However, the benefits provided by Fannie Mae and Freddie Mac are broad-based, encouraging more housing construction and consumption throughout the income and social spectrum. T]he bulk of their mortgage purchases do not involve the group that ought to be the target of ownership-encouraging activities. Consistent with this, it appears that their activities have not appreciably affected the rate of home ownership in the United States. [Instead, they have encouraged people who already have homes, or who would have bought anyway, to buy more or larger homes. As a result of this and other broad-based inducements to home ownership, t]he United States has too much housing (at the expense of other goods and services), and Fannie Mae and Freddie Mac make it worse (while not doing an especially good job of focusing on the low- and moderate-income first-time buyer where the social argument is strongest).
Safety and Soundness

To the extent that the financial markets are correct in their belief about the implicit guarantee—that the U.S. government would come to the rescue of their creditors if either of the two companies experienced financial difficulties—a moral hazard problem is created: The creditors do not monitor the two companies’ managements as closely as they would if the creditors were more fearful of losses. In turn, the managements can engage in activities that involve greater risk, since the companies’ owners will benefit from the “upside” outcomes while (because of the protections of limited liability) being buffered from the full consequences of large “downside” outcomes. The creditors’ guarantor—the federal government—is thus exposed to potential loss.

This problem of moral hazard is a general problem for the creditors of a limited liability corporation. Outside of the financial sector, creditors long ago realized the existence of the problem and created monitoring structures, as well as restrictions in lending agreements and covenants in bonds, that give creditors the ability to restrain owners’ and managers’ risk-taking, especially when net worth levels diminish. For banks and other depositories, where the institution’s primary creditors are considered to be less capable of monitoring and protecting themselves against this moral hazard behavior and where the consequences of bank insolvency failures have been considered economically serious (e.g., the potential problem of contangion) and politically serious (the losses experienced by individual depositors), federal and state safety-and-soundness regulation has been the public-sector substitute for the private monitoring just described. The federal government’s direct exposure to losses, because of federal deposit insurance (since 1933)[,] provides another justification for such regulation.

With respect to Fannie Mae and Freddie Mac, the federal government’s exposure to potential losses from excessive risk taking or even just from errors and poor judgments would logically call for safety-and-soundness regulation, akin to that applied to banks. Only in 1992, however, did Congress come to that realization, in the Federal Housing Enterprises Financial Safety and Soundness Act. The act created the Office of Federal Housing Enterprise Oversight, lodged within HUD, as the safety-and-soundness regulator for the two companies and instructed the agency to develop forward-looking risk-based capital requirements for them. Only 10 years later did the agency succeed in issuing a final set of those rules. That delay, plus Fannie Mae’s revelation of a large exposure to interest-rate risk in 2002 and Freddie Mac’s revelation in 2003 of the necessity for a massive restatement of its recent years’ income and balance sheet statements, have led to calls for strengthening the regulatory structure. . . . As of September 2004, no definitive legislative action had been taken.

Systemic Risk

The general notion of systemic risk is that the financial problems of one institution could have wider spread effects on other parts of the economy. For commercial banks, a “contagion” effect is one such scenario, whereby depositor “runs” on one shaky bank might cause worried depositors of other banks to withdraw their cash from those banks, which would create a liquidity squeeze for those latter banks; or the liquidation of assets

428
Draft — do not distribute!
by the banks in their efforts to meet their depositors’ claims could depress asset values sufficiently so that other banks’ asset values and solvency were impaired. Alternatively, there might be a “cascade” effect, whereby the chain of banks’ claims on one another would mean that the insolvency of one bank would reduce the asset values of other banks that had claims on the first bank (and this cascade could lead to and reinforce a contagion problem, and vice-versa).

. . . Fannie Mae and Freddie Mac face two major categories of risk: credit risk (i.e., the risk that mortgage borrowers will default on their payment obligations and that the prices of the repossessed housing are below the outstanding loan balances, which would impair the value of the mortgage assets in the companies’ portfolios and/or require the companies to make payments on their MBS guarantees); and interest-rate risk (i.e., the risk that interest rates change after the investment in a mortgage, and the risk that changes in interest rates could cause the values of their mortgage portfolios to fall below the values of their outstanding debt obligations).

There is general agreement that the credit risk on most single-family residential mortgages has been quite low. . . . [But if] there were to be a Great Depression-type of collapse in housing values, however, these credit-risk losses could deteriorate considerably.

Instead, the focus has been on interest-rate risk—on the risk that interest rates may change, which would affect the market values of Fannie Mae and Freddie Mac’s mortgage assets and MBS. This concern, of course, applies only to the assets held in the portfolios of the two companies, since the holders of their MBS are the bearers of the interest-rate risk on those MBS. By holding large portfolios of largely long-term fixed-rate mortgages and MBS that can be prepaid without penalty, the two companies potentially exposed themselves to extensive interest-rate risk. . . .

Any discussion of the systemic consequences must start with the sheer sizes of the two companies: Their portfolio holdings and outstanding MBS now account for almost half of the total of all residential mortgages. On the one hand, this size is a potential element for stability: At times of externally generated stress (e.g., the market stress of September 11, 2001; the potential market meltdown related to the demise of Long Term Capital Management in September 1998; or the stock market free-fall of October 1987), their continued participation in the secondary mortgage markets has been and can continue to be a source of strength and stability for those markets. If either of them were to begin to falter financially, however, then their size would become a systemic liability. . . .

With respect to effects on existing liability holders, systemic effects (beyond just the direct losses experienced by the liability holders and counterparties) would depend on the extent of the direct losses and the extent to which the directly exposed parties are themselves leveraged (and thus their losses can impose further losses on others). The extent of a GSE’s losses in the event of financial difficulties is difficult to predict. . . .

With respect to a contagion or cascading effect of creditor losses, the primary candidates would be depository institutions, which (in aggregate) hold about a sixth of the two
companies’ debt and about 40 percent of their MBS and which are allowed by regulation to hold unlimited amounts of their obligations. A recent study shows that, as of the third quarter of 2003, depositories’ aggregate holdings of the two companies’ debt came to 3.3 percent of all depositories’ assets, or slightly more than a third of their aggregate net worth (which was about 9.1 percent of assets), while their aggregate holdings of the GSEs’ MBS came to 8.5 percent of their aggregate assets. Though losses of value of GSE debt and MBS of, say, 5 percent would be far from a welcome event for depositories, it would also be far from a devastating event for most of them and would be unlikely to have widespread systemic consequences.

As for the direct effects on mortgage markets of financial difficulties by one of the companies, it is difficult to imagine that there would be no consequences when an $800 billion or $1 trillion company withdraws from its primary activities. But the extent of the consequences would depend on whether and to what extent and how quickly the other GSE could pick up the slack, as well as how elastic would be the responses of the other major providers of residential mortgage finance. Since no such event has occurred, it is difficult to provide estimates of magnitudes.

\[The Absence of Prepay Penalties and the Bearing of Interest-Rate Risk\]

The standard residential mortgage in the U.S. is a long-lived, fixed-rate debt instrument, which the borrower can prepay at any time with no penalty. Fannie Mae and Freddie Mac are both cause and effect with respect to these characteristics, since over 90 percent of the mortgages that they buy are fixed-rate instruments, and they rarely buy mortgages that have prepay penalties. The absence of prepay penalties exacerbates the interest-rate risk that is borne by the holder of a mortgage or MBS.

[Consider a nonprepayable mortgage. The value of this mortgage to the lender increases when interest rates fall and decreases when interest rates rise. But now allow the borrower to prepay his mortgage. He will be more likely to prepay when rates fall, either by refinancing or by selling his home and buying a new one; and the lender will only be able to reloan those funds at the new, lower rate. So with prepayment, a lender is less able to benefit from falling rates, but he is still stuck with the adverse consequences of rising rates. Thus, the borrower’s ability to prepay, by removing part of the upside of interest-rate fluctuations, exacerbates the lender’s interest-rate risk.]

\[This extra risk borne by the lender is not free to the borrower. Instead, the risk of the borrower’s exercising her option to prepay without penalty is incorporated into the overall interest rate and fees that a competitive market will charge all borrowers (so long as the lender cannot determine beforehand which borrower is more likely to prepay). Accordingly, even those borrowers who (for whatever reason) are unlikely to prepay their mortgages must pay extra because of the greater risk imposed on lenders, and there is a cross-subsidy that runs from those who are less likely to prepay to those who are more likely to prepay.\]

Why is the prepay option not priced more explicitly—for example, through an explicit penalty for prepaying (which, in turn, would allow a lower interest rate and lower initial
fees)? Or, at least, why are borrowers not more often offered the choice between a no-prepayment-penalty mortgage (with a higher interest rate and initial fees) and a prepayment penalty mortgage (with a lower interest rate and initial fees)? At least part of the reason appears to be a patchwork of state regulations that, in some states, limits (or forbids) the ability of state-chartered depositories to impose prepayment penalties. However, the buying patterns of Fannie Mae and Freddie Mac—they almost exclusively purchase no-prepayment-penalty mortgages[, perhaps because believe they have a comparative advantage at managing this exacerbated interest-rate risk]—also influence the outcome.

As a related matter, so long as the lending/borrowing arrangement with respect to a home involves long-term finance, interest-rate risk unavoidably arises and cannot (from an economywide perspective) be diversified away. Two questions then arise: (1) Who bears the interest-rate risk? and (2) Who should bear that risk?

The first question is easier to answer: With adjustable-rate mortgages (ARMs), the borrower bears the risk. With long-term fixed-rate mortgages that are nonprepayable, the lender or the MBS holder bears most of the risk. The interest-rate risk borne by lenders on long-term fixed-rate mortgages is exacerbated by the current practice of allowing borrowers to prepay their mortgages without penalty. In essence, the penalty-free prepay option allows the borrower to shed all interest-rate risk.

The second question is harder. Some general principles can be stated, however. First, diversification of that risk is surely a good thing. Second, the bearing of that risk should be by individuals or institutions that are knowledgeable and skilled at managing the risk and that are in a financial position to bear it without undue financial hardship (and without creating the transactions costs of bankruptcies, etc.). This surely argues for allowing (but not requiring, nor forbidding) mortgage originators to offer ARMs to those knowledgeable borrowers who want them. It also argues for allowing mortgage originators to offer fixed-rate mortgages that may (or may not) include prepayment penalties, which would allow the prepayment risk to be explicitly priced, and then letting market participants choose. It is far from clear that the federal government needs to be the explicit or implicit backstop for this process through its maintenance of the special GSE status of Fannie Mae and Freddie Mac (or of the FHLB system).

Efficient Transmission of Benefits

Within the sphere of conforming mortgages, Fannie Mae and Freddie Mac are a protected duopoly, which could affect their pricing behavior and thus the extent to which they pass through to homebuyers the benefits that they receive as a consequence of their special GSE status.

It does appear that the two companies have held on to at least part of their special benefits and have earned supranormal returns. There is a deep irony to the consequences of this exercise of market power for allocative efficiency: To the extent that one believes (as was argued above) that public policies in general encourage too much housing and that the two companies’ activities make it worse, their exercise of market power
(implying that mortgage rates are not as low as if they behaved wholly competitively and thus home buying is not as encouraged) means that global allocative efficiency is improved.

Possible Inherent Efficiencies or Inefficiencies

. . . Skepticism is warranted as to whether the two companies’ special GSE status adds extra efficiency to mortgage markets. First, as is argued above, the current broad-based approach of the two companies is surely not an efficient way to address the positive externality with respect to home ownership.

Second, there is the issue of transactions costs with respect to the credit risk on residential mortgages. To provide assurances of timely payments to the holders of their MBS, “private-label” (i.e., private, non-GSE) issuers usually create a senior-subordinated structure (whereby the subordinated security absorbs the first credit losses) that protects the holders of the senior MBS. The creation of this structure involves transactions costs, as does the process of obtaining a rating on the senior MBS from one or more rating agencies (e.g., Moody’s or Standard & Poor’s). Investors in the senior MBS would be interested in learning whether some of the subordinated MBS tranches are experiencing greater losses than was expected (which would mean greater risks for the associated senior MBS), thus entailing further monitoring (transaction) costs. The blanket credit-loss guarantees issued by Fannie Mae and Freddie Mac (backed by the implicit federal guarantee) eliminates all of those costs.

This argument is surely correct as far as it goes. But the Fannie-Freddie process guarantee must then be backstopped by the government-oriented safety-and-soundness regulatory process described above and ultimately by taxpayers. Which route offers the lowest costs (short run or long run) is not obvious.

A third argument is that the new-era securitization process is inherently more efficient than is the depository-driven process of yore and that the expansion of Fannie Mae and Freddie Mac (with their implicit federal guarantee) at the expense of depositories’ (with their explicit federal deposit insurance) holdings of conforming residential mortgages is evidence of this superior efficiency. However, regulatory considerations have also played an important role in the GSEs’ growth. Commercial banks and S&Ls have been encouraged to hold MBS rather than whole mortgage loans by risk-based capital requirements (which have been in place since 1988) that require only 1.6 percent capital for holding any MBS that is rated AA or better but require 4 percent for holding unsecuritized residential mortgages. Further, the capital requirements that have applied to Fannie Mae’s and Freddie Mac’s holdings of mortgages (2.5 percent) have been substantially less than the capital requirements that have applied to depositories’ holding of mortgages (4 percent), giving the former a cost advantage. Accordingly, though mortgage securitization (as an efficient innovation of the 1970s and 1980s) was surely going to grow and gain market share relative to the traditional depository route, the extent of the GSEs’ growth is not necessarily an indicator of their special and inherent efficiencies.
As for possible inherent inefficiencies that may accompany the two companies’ special
GSE status, there is no assurance that the current managements and organizational
structures for Fannie Mae and Freddie Mac are the most efficient for doing what they do.
Since Congress has issued only two charters of this particular kind, the ability of
competitive processes to reward more efficient firms and winnow less efficient firms
from the marketplace is inhibited. Further, the two firms are not required periodically to
bid for their franchises in an auction against potential replacements; they have been
“grandfathered” indefinitely. And the market for corporate control cannot operate
effectively: Their limited charters make them immune to takeover by any other firm, and
their large size and special GSE status make them virtually immune to a “hostile”
takeover by an investor group.

As a related matter, whenever either of the two firms has expanded slightly in
“horizontal” (e.g., subprime lending) or “vertical” (e.g., providing underwriting software
to mortgage originators) direction—or even publicly contemplated such moves—critics
have complained that the two companies’ ability to expand arises solely from the low-
cost funding that they enjoy because of the implicit guarantee and not because of any
inherent efficiency advantage (and that they are in fact elbowing aside inherently more
efficient enterprises). Without a “clean” market test, there is no way to resolve such
questions.

**What Is to Be Done?**

*First-Best Option*

The analysis provided above points in a clear direction with respect to Fannie Mae and
Freddie Mac: Since there seems to be no special efficiency reason for preserving their
special GSE structure, since they mostly just add to an already excessive amount of
encouragement for housing in the United States, since their role in addressing the
important social externality of home ownership is modest at best, and since the implied
guarantee (to the extent that it would be honored) creates a contingent liability for the
U.S. government, an outright privatization of the two companies—the withdrawal of their
special charters and their conversion to normal corporations—would be the first-best
outcome. This would imply that the two companies would no longer enjoy any special
privileges, but would no longer be restricted to their current narrow slice of the financial
world. How these companies and their owners would fare in that scenario would then be
a matter for markets, and not the Congress or OFHEO, to decide....

The consequence of true privatization for residential mortgage markets would be modest:
Mortgage rates would be about 25 basis points higher than would otherwise be the case.
Grass would surely not grow in the streets of America as a consequence—and would
surely continue to grow in most backyards. And, because the United States already builds
and consumes too much housing, this would be a move in the right direction.

In their place, the federal government ought to deal directly with the true positive
externality related to housing: encouraging low- and moderate-income first-time buyers.
Such a program should be an explicit on-budget encouragement for such home purchases, with subsidies for down payments and for monthly payments.

As part of this true privatization, the secretary of the Treasury should state clearly at the congressional hearings that the Treasury (after the passage of the privatization legislation) would treat the two companies just like other corporations in the U.S. economy, would not consider the two companies to be “too big to fail,” and would have no intention of “bailing them out” in the event of subsequent financial difficulties. The president should reiterate this message at the official signing of the legislation. Also, bank and S&L regulators should revise their “loans-to-one-borrower” regulations so that depositories’ holdings of Fannie Mae and Freddie Mac debt would be treated similarly to their holdings of other companies’ debt (i.e., loans to any single borrower normally cannot exceed 10 percent of the depository’s capital), rather than the unlimited holdings that are currently permitted.

Further, in order to ameliorate the concentration of interest-rate risk that the current structure of fixed-rate mortgages without prepayment penalties places on lenders or MBS holders and that may be an extra element that unduly strengthens the role of Fannie Mae and Freddie Mac in the mortgage markets, lenders should have the freedom to offer mortgages that would include a fee for the prepayment option that is usually not explicitly priced (but is surely included in the overall pricing of mortgages). Such explicit pricing will also eliminate the cross-subsidy that currently runs from those who do not exercise the option to those who do. State laws and regulations that inhibit such explicit pricing should be repealed. . . . [The authors also suggest ways to reduce the cost of housing.]

Second-Best Measures

The true privatization of the two companies may well be unlikely in the current political environment. The political attractiveness of an arrangement that reduces housing costs but has no on-budget consequences is powerful. Accordingly, second-best measures should be considered.

First, regardless of what’s done with respect to Fannie Mae and Freddie Mac, an explicit housing program for low- and moderate-income first-time buyers is worth undertaking in its own right. So are any efforts to allow explicit pricing of the prepayment option and the efficiency-enhancing efforts to reduce the cost of housing.

Second, even if the two companies retain their GSE status, bank and S&L regulators could still apply the loans-to-one-borrower limitations to depositories’ holdings of their debt, as suggested above.

Third, as a way to reduce the financial markets’ belief in the “implicit guarantee,” the secretary of the Treasury should state loudly and at frequent intervals that it is the policy of the federal government to adhere to what is stated on the Fannie Mae and Freddie Mac securities: that these are not obligations of the federal government and that the Treasury

Draft — do not distribute!
has no intention of “bailing them out” in the event that they become financially troubled. As discussed above, such explicit denials have not been enunciated in the past.

Fourth, in addition to keeping or even increasing the pressures of HUD’s affordable housing “mission” goals with respect to the two companies’ purchases of mortgages, the two companies should be forced to concentrate further on the lower end of the housing market by freezing the conforming loan limit at its current level of $333,700 and waiting for median sales prices (or 80 percent of median) to catch up to that level before resuming indexed annual increases. This freeze would also have the beneficial effect of limiting the two companies’ growth and thereby reducing potential systemic risks.

Fifth, the safety-and-soundness regime should be strengthened through the transfer of OFHEO to the aegis of the Treasury, with a structure and powers (especially receivership powers) that resemble those of the regulatory agencies for depository institutions that are currently housed within the Treasury: the Office of the Comptroller of the Currency (for national banks) and the Office of Thrift Supervision (for S&Ls and savings institutions). The major argument against such strengthening is, as was discussed above, the risk that such strengthening would also strengthen the financial markets’ belief in the implicit guarantee. Though this possibility is troubling, the dangers of not strengthening the regulatory regime appear to be even greater. . . .


Executive Summary

Leading political figures . . . are proposing to offer a new public insurance option to Americans who lack employment-based coverage. The public plan would be similar to conventional Medicare (the “public Medicare plan,” as distinguished from private plans that contract with Medicare) in that it would be managed by the federal government and pay private providers to deliver care. The public plan would be offered through a new national insurance “exchange,” where it would compete with private insurance plans.

This policy brief sets out the argument for public plan choice. The core argument is that public insurance has distinct strengths and thus, offered as a choice on a level playing field with private plans, can serve as an important benchmark for private insurance within a reformed health care framework. This is not an argument for a universal Medicare program, but instead for a “hybrid” approach that builds on the best elements of the present system—large group plans in the public and private sectors—while putting in
place a new means by which those without access to secure workplace insurance can choose among health plans that provide strong guarantees of quality, affordable coverage. The case made in this brief is that this menu of health plans must include a good public plan modeled after Medicare if the broad goals of reform—universal insurance and improved value—are to be achieved.

First, public insurance has a better track record than private insurance when it comes to reining in costs while preserving access. By way of illustration, between 1997 and 2006, health spending per enrollee (for comparable benefits) grew at 4.6 percent a year under Medicare, compared with 7.3 percent a year under private health insurance. At the same time, Medicare has maintained high levels of provider participation and patient access to care.

Medicare has proven superior at cost control not just to health plans in the private sector, but also to private plans that contract with the federal government, such as those offered through the Federal Employees Health Benefits Program (FEHBP)—suggesting that public insurance can outperform private plans even in the context of insurance reforms.

Second, over the last generation, public insurance has pioneered new payment and quality-improvement methods that have frequently set the standard for private plans. More important, it has the potential to carry out these vital tasks much more effectively in the future, using information technology, large databases of practices and outcomes, and new payment approaches and care-coordination strategies. Indeed, a new public plan could spearhead improvement of existing public programs as well as private plans.

Third, public plan choice is essential to set a standard against which private plans must compete. Without a public plan competing with private plans, we will continue to lack strong mechanisms to rein in costs and drive value down the road. As a benchmark, a new public plan alongside private plans will help unite the public around the principle of broadly shared risk while building greater confidence in government over the long term.

Public plan choice will allow Americans to realize the benefits of both public and private plans: flexibility and security, innovation and stability, and market and democratic accountability. . . .

**What Is Public Plan Choice?**

In essence, public plan choice is simple. Many reform plans envision the creation of new national or regional purchasing pools. Such pools, often called “exchanges,” allow those without good employment-based insurance to choose among large private health plans, providing a basis for group pooling of medical risks similar to that provided by large employers. The public insurance option simply makes a public plan available alongside the private plans that can be enrolled in through the exchange. This plan will compete with private plans, ensuring an insurance product with broad choice of providers and encouraging private plans to match the administrative efficiencies, cost-control abilities, and quality-improvement capacities of public insurance.
Although simple in broad conception, public plan choice raises a number of questions about how the public plan should be structured (whether, for example, it would piggyback on Medicare or be separate), how it should compete with private plans (how, for example, the private plans would be paid to ensure that they did not cherry pick healthy patients, leaving less healthy enrollees in the public plan), and who should be able to enroll in the plan (whether, for example, it would replace Medicaid for working poor Americans or only be available to workers without other sources of coverage). The last section of this brief takes up the most important of these issues, namely, how to create a level playing field for public-private competition. The main purpose of the brief, however, is to discuss the rationale for public plan choice, not the exact form it should assume.

By the same token, the brief does not take up another attractive idea that is compatible with public plan choice: opening up Medicare to 55 to 64-year olds without workplace coverage. . . . Nor, finally, does this brief discuss proposals for simply expanding Medicare coverage . . . . Instead, the analysis focuses on proposals with two overarching features: (1) they include some sort of “pay or play” requirement permitting employers to choose between contributing to the cost of covering their workers through a new insurance exchange or providing coverage directly, and (2) they allow those enrolled in the exchange to choose between public and private insurance—that is, public plan choice. . . .

**Cost-Control Advantages of Public Insurance**

It is often assumed that private health plans are much more efficient than public health insurance. Yet a range of studies demonstrate that public insurance is able to provide a given level of benefits for less than they would cost through private insurance. Lower administrative costs and the ability to bargain for lower service and drug prices chiefly explain this advantage, as does the obvious lack of a profit margin in public programs. These features of public insurance not only allow it to offer the same coverage for less than private plans. They also, the evidence suggests, allow it to better restrain the increase in costs over time while preserving inclusive coverage.

The remainder of this section focuses on the relative performance of Medicare and private health insurance in controlling costs. The Medicare program is under financial strain and has evident flaws that require correction, but it has performed far better relative to private health insurance than conventional wisdom suggests. And, as the next section of this brief discusses, a new public plan modeled on Medicare could do even better.

**Administrative Efficiencies**

Perhaps the most obvious advantage of public insurance is that it is inexpensive to administer. The public Medicare plan’s administrative overhead costs (in the range of 3 percent) are well below the overhead costs of large companies that are self-insured (5 to 10 percent of premiums), companies in the small group market (25 to 27 percent of premiums), and individual insurance (40 percent of premiums).
These administrative spending numbers have been challenged on the grounds that they exclude some aspects of Medicare’s administrative costs, such as the expenses of collecting Medicare premiums and payroll taxes, and because Medicare’s larger average claims because of its older enrollees make its administrative costs look smaller relative to private plan costs than they really are. However, the Congressional Budget Office (CBO) has found that administrative costs under the public Medicare plan are less than 2 percent of expenditures, compared with approximately 11 percent of spending by private plans under Medicare Advantage. This is a near perfect “apples to apples” comparison of administrative costs, because the public Medicare plan and Medicare Advantage plans are operating under similar rules and treating the same population.

... The experience of private plans within FEHBP carries the same conclusion. Under FEHBP, the administrative costs of Preferred Provider Organizations (PPOs) average 7 percent, not counting the costs of federal agencies to administer enrollment of employees. Health Maintenance Organizations (HMOs) participating in FEHBP have administrative costs of 10 to 12 percent.

**Bargaining Leverage**

The government has another advantage when it comes to holding down costs: It is capable of using its concentrated purchasing power to pioneer new payment methods that bring down costs. Medicare’s improving cost-control performance over the last quarter century tracks closely the introduction of innovative changes in hospital payment using a prospective payment system in 1983 (a system by which hospitals are paid a pre-determined rate for each Medicare admission based on the patient’s diagnosis at the time of admission) and the creation of a resource-based physician fee schedule (a scale of national uniform relative values for all physicians’ services) and volume controls on overall Medicare physician spending in the 1990s. While Medicare’s methods of paying providers clearly require improvement, especially with regard to physician payment, the program’s record is still notably superior to that of private insurance.

Perhaps the simplest way to look at Medicare’s bargaining power is to compare Medicare rates with those paid by private insurance. According to the Medicare Payment Advisory Commission (MedPAC), Medicare’s rates for physicians are 81 percent of private rates—a clear sign of superior negotiating leverage. For hospitals, MedPAC estimates that Medicare pays around three-quarters of what private payers do. These differentials have been relatively stable, and as noted below, they have not had the negative effect on provider participation or revenues that critics often suggest. Indeed, for-profit hospitals made record profits in 2007, and the number of physicians billing Medicare is actually increasing faster than enrollment in Part B medical insurance.

Another source of comparative insight is the relative costs of the public Medicare plan and private plans that contract with the program through Medicare Advantage.

As is well known, Medicare Advantage plans are substantially overpaid relative to what it would have cost to provide coverage to the same enrollee in the public Medicare plan—13 percent more on average per person. This overpayment reflects two main
problems: a method for paying plans that subsidizes their participation in Medicare Advantage and the ability of the plans to attract healthier (and hence less costly) people with Medicare. Both of these problems can and should be addressed—in Medicare and in any new framework for public-private competition.

Yet the larger lesson of Medicare Advantage is that private plans do not appear to have strong tools for controlling costs relative to the public Medicare plan. The most tightly regulated HMOs have been shown to perform roughly as efficiently as the public Medicare plan does, but according to MedPAC, most private plans are not as efficient as the public Medicare plan. All but HMOs bid to provide Medicare Part A and Part B benefits for more than the public Medicare plan spends on the same benefits—often much more. Indeed, the fastest-growing category of Medicare Advantage plan, private fee-for-service plans[,] are the least efficient and most costly for Medicare, with their bids for Part A and B benefits fully 108 percent of the public Medicare plan’s costs.

Were Medicare permitted to bargain directly for drug prices, moreover, there is no question it would receive better deals than currently offered to private payers. The CBO has found that drug prices under four federal programs—including the Veterans Health Administration (VHA) and Medicaid—are on average 49 percent below the average wholesale price of the drugs. Another recent study found that the lowest price available for the top 20 drugs prescribed to seniors were 58 percent cheaper under the VHA plan than under Medicare Part D. Medicare’s private plans negotiated drug manufacturer rebates of only 8.1 percent in 2007.

The failure of private insurers to obtain affordable prices is borne out by international comparisons as well. A recent McKinsey study finds that branded drugs in the United States are 60 percent more expensive than in Canada, with its “single payer” provincial insurers, and that the top-selling drugs of leading drug companies are 2.3 times more expensive here than in other rich nations, where public-sector bargaining is prevalent.

Is Bargaining Unfair?

Government’s use of its countervailing power to hold down prices is often criticized. A recent study released by [hospital and insurance groups] claims, for instance, that Medicare and Medicaid grossly “underpay” providers, leaving private insurers to pick up the difference. The study simply assumes, however, that all payers should pay the same rates and that the total level of payments to providers is appropriate. . . . The whole point of bargaining, however, is to gain volume discounts and restrain total spending—insofar as doing so is consistent with ensuring good access to providers and high-quality care. So far, there is little evidence that Medicare bargaining has undermined access or quality.

It is worth remembering, after all, that price bargaining is exactly what HMOs and other big health plans were supposed to do—only Medicare appears to do it better. The consolidation of the private insurance market over the last two decades was widely expected to bring down costs. (In 16 states, the dominant carrier accounts for at least 50 percent of private enrollment; in 36 states, the top three carriers account for at least 65 percent of the market.) Yet it obviously has not. Instead, private plans are passing on
rising costs to individuals while increasing their profitability. The reasons for this are multiple, and they go to the heart of the argument for a public plan alongside private plans.

First, the hospital market has grown increasingly concentrated, giving providers considerable market power of their own in negotiations with insurers. . . . Second, private insurers appear to have largely acquiesced to these price increases. . . . Both of these trends provide strong reason for doubting that private insurance payments are the appropriate standard for public payments. . . .

Nonetheless, the effects of concentrated purchasing power on the revenues of providers is understandably of central concern. Studies of cost shifting by the public Medicare plan onto private payers have produced mixed results: The general conclusion is that there is some, albeit a much smaller amount than suggested by critics of Medicare pricing. A careful 2006 study of hospital cost-shifting in Medicare concludes that “a 1 percent relative decrease in the average Medicare price is associated with a 0.17 percent increase in the corresponding price paid by privately insured patients”—meaning that around 17 cents of every dollar in relative reductions in public Medicare plan payments to private hospitals are shifted onto private patients. If this estimate is correct, then cost shifting from the public Medicare plan amounted to less than 10 percent of the overall increase in hospital prices to private payers between 1997 and 2001—the period under study.

[MedPAC’s report suggests that hospital services are not becoming scarcer, and that Medicare recipients are not having trouble finding doctors.]

Finally, it is worth emphasizing that any national reform proposal that included public plan choice would involve a major expansion of insurance coverage. (Many reform proposals would also upgrade Medicaid payment rates to bring them closer to Medicare and private payments.) This expanded coverage would mean that providers would be paid for a much higher share of the services they delivered. These new revenues could well exceed any negative income effects that occurred due to a public plan bargaining for lower prices.

Long-Term Cost Control

The evidence strongly indicates that a public plan can provide coverage less expensively than private insurance without impairing access or quality. Yet the greatest potential cost-control advantage of a public plan is its ability to restrain the rate of increase of costs over time—the key to maintaining good coverage without excessively burdening public and private budgets.

Although you would not know it from the debate over Medicare’s finances, Medicare has become increasingly effective at restraining the “excess growth” of spending—that is, per capita cost growth in excess of overall economic growth (accounting for population aging). Excess cost growth is a critical measure of the sustainability of any health plan, because it shows how quickly spending will rise as a share of personal and business budgets over time. A recent study that examined excess growth in spending on Medicare
services between 1975 and 2005 found that the annual rate of excess growth fell from 5.6 percent during 1975–1983, to 2.1 percent during 1983–1997, to 0.5 percent during 1997–2005. Put another way, since Medicare payment controls were put in place in the early 1980s, Medicare spending has grown much more slowly than in the past—and in the most recent period (1997 to 2005), it grew only slightly faster than the economy overall, adjusting for population aging.

It is not possible using these data to compare Medicare excess growth directly with private insurance excess growth. However, a rough comparison is provided by contrasting excess spending growth among the nonelderly, most of whom are covered by private insurance, with excess cost growth among the elderly, 97 percent of whom are covered by Medicare. . . . [E]xcess cost growth for the nonelderly was 3.4 percent between 1996 and 2004. That is, spending grew 3.4 percentage points faster than the economy. The comparable figure for the elderly—again, virtually all of whom are covered by Medicare—was 0.3 percent.

A more direct comparison is provided by examining Medicare and private insurance spending for comparable benefits in recent decades. . . . [P]rivate plans’ spending per enrollee has grown substantially faster than Medicare spending per enrollee, especially in the last decade or so. Private insurance outlays per enrollee grew an average of 7.6 percent a year between 1983 and 2006, compared with 5.9 percent growth in per enrollee spending under Medicare—a 22 percent difference. . . . The gap is even bigger in recent years. . . . [N]ot only has Medicare more successfully restrained the rate of increase of per enrollee spending, the rate of growth is also on a steeper downward trajectory under Medicare than under private insurance.

The Federal Employees Health Benefits Program has frequently been invoked as a model for national reform, and indeed it provides one template for a national insurance exchange offering competing private plans. It is not, however, a model of cost restraint when compared with Medicare. . . . FEHBP’s annual growth rate of per enrollee spending averaged 7.3 percent from 1985 to 2002 . . . compared with 5.8 percent for Medicare. Indeed, the growth rate for FEHBP is virtually identical to that for private health insurance over this period . . . . This suggests that simply replicating FEHBP on a broader scale—without public plan choice—would be unlikely to provide the long-term cost restraint essential for successful reform.

A similar story is told by foreign experience. Other rich nations all rely on public or quasi-public insurance more than the United States does. (They do not, however, all rely on a single public insurer, and many have public and private insurance operating side by side.) And taken as a whole these nations not only spend much less on health care as a share of their economy than we do; they have seen their health costs slow dramatically in recent decades, while U.S. costs have continued to grow much faster than the economy. . . .
A PUBLIC PLAN CAN SPEARHEAD QUALITY IMPROVEMENTS

The Surprising Success Story of the VHA

. . . Perhaps the most powerful example of how investments in quality improvement by a public plan can pay off is provided by the Veterans Health Administration. The VHA has used its integrated framework to create a model evidence-based quality-improvement program that delivers the highest quality care in the nation, as measured by adherence to established treatment protocols. In the rest of American health care, only around half of adults and children receive the care they should. The share in the VHA is over two thirds.

. . . Beginning in the early 1990s, VHA leadership instituted both a sophisticated electronic medical record system and a quality measurement approach that holds regional managers accountable for several processes in preventive care and in the management of common chronic conditions. Other changes include a system-wide commitment to quality improvement principles and a partnership between researchers and managers for quality improvement.

The VHA’s promulgation of specific performance measures and emphasis on accountability—possible only because of the broad reach of its coverage—appear to be at the heart of its success. The use of computerized reminders and electronic records; the emphasis on standing orders, improved inter-provider communication, facility performance profiling, leveraging of academic affiliations, and accountability of regional managers for performance; and the creation of a more coordinated delivery system—in tandem, these reforms have allowed the VHA to create and uphold high standards of quality.

Medicare’s Improving Quality Record

The Medicare program is not, of course, the VHA, and some of the lessons provided by the VHA integrated system are not applicable to an insurance program like Medicare. Yet key elements of the VHA strategy—notably, greater emphasis on research-based coverage decisions, improved use of information technology, and increased stress on performance measures and accountability—could be effectively used in Medicare and a new public plan for the nonelderly, and indeed it is unclear how they could be developed without such a coordinated public-sector effort.

Medicare already shows unique quality advantages over private insurance when it comes to reliable patient access to affordable care—advantages that would carry over to a new public plan for the nonelderly. Elderly Americans with Medicare report that they have greater access to physicians for routine care and in cases of injury or illness than do the privately insured. They are also half as likely as nonelderly Americans with employment-based insurance to report common access problems, such as skipping a medical test, treatment, or follow up, and failing to see a doctor when sick.

Over the last two decades, moreover, Medicare has increasingly emphasized improved payment methods and rigorous reviews of technology and treatment, and it has made increasing investments in quality monitoring and improvement. Revealingly, private
plans generally use the public Medicare plan’s criteria for covering treatments as their standard of medical necessity, and they have adopted many of Medicare’s innovations in payment methods. . . . “Traditional Medicare has been the source of important payment innovations, moving many payment systems away from fee-for-service to prospective payment, such as the diagnosis-related group (DRG) prospective payment system (PPS) for inpatient services. The resource-based relative value scale (RBRVS) for physician fees, despite its flaws, has been adopted widely by private plans . . . . Commercial insurers also look to Medicare to make initial technology approval decisions and to initiate more-aggressive payment denials—for example, for ‘never’ events and medically ineffective treatments.” . . .

Still, much more needs to be done. [Recommended Medicare innovations, which could be adopted by a public plan and made available to private payers,] include:

- Developing practice guidelines and quality measurements that will allow for value-based purchasing (a policy mechanism that links payment to performance).
- Requiring public reporting by providers of quality indicators to help purchasers and payers get maximum value.
- Testing the effectiveness of new technology.
- Developing a pay-for-performance system based on quality outcomes.
- Finding alternatives to the fee-for-service-based system for physician payment.
- Shifting payment methods and rates to better reward primary care providers and increase their supply and to decrease the oversupply of specialty physicians, who are escalating costs without necessarily improving quality.
- Building a system based on coordinated care for those with chronic diseases, rather than maintaining our current fragmented care.
- Removing wide geographic variations in care from one region of the country to another, which are largely driven by a community’s supply of specialists and technology rather than the services patients actually need.

Public Insurance Has the Potential to Lead the Quality Revolution

Medicare and a public plan for those under 65 would be well positioned to lead these efforts—if they were given the tools to collect and maintain extensive outcomes data, test new methods of providing and paying for care, and use their market power to promote quality and cost effectiveness in both the public and private sectors.

The simple truth is that private insurance has few incentives to conduct comparative-effectiveness research, and limited scope to influence the practices of providers and other insurers even when they do. Comparing the clinical effectiveness of tests, procedures, and drugs with their alternatives is critical to increasing effectiveness and reducing costs. But for insurance companies, it is expensive and the benefits, if made public, are not theirs alone. As MedPAC has noted, “Because the [public dissemination of] information can benefit all users and is a public good, it is underproduced by the private sector.”

Moreover, insurance companies are generally reluctant to share private information that will allow others to learn lessons about how best to contain costs and improve quality. . . .

443
Draft — do not distribute!
U.S. News and World Report recently noted that 126 health care plans refused to provide data to a national accrediting agency that was needed for the magazine to rank plan performance. Transparency in Medicare has helped identify huge variations in spending per capita across the country and to determine that areas with higher per capita spending scores no better on quality measures, and often score worse. Private insurers have far fewer incentives to make such information public.

In addition, private insurers have limited incentive to attract or treat those with chronic and costly disease or behavioral health problems such as obesity—the patients who are least likely to sign up for private plans in Medicare Advantage. This is the population that private insurers have the greatest incentive to avoid through targeted advertising and risk selection (which, existing research suggests, can only be partially addressed through better risk adjustment). Plans that adopt innovative strategies for disease management may find themselves attracting less healthy patients, discouraging them from engaging in such innovations. On the other side of the equation, patients with greater health needs may desire the greater stability or choice of providers that a public plan can provide. . . . [T]he public Medicare plan attracts people with poorer health status than do Medicare private plans. Yet these are precisely the patients most in need of innovations in treatment and care coordination. A public plan, which by nature will take all comers and will be attractive to those in poorer health status, is best poised to improve the treatment of these patients and disseminate the lessons learned to the private sector.

Finally, participation in public plans is much more stable. Insurers move in and out of markets, change their benefits frequently, shift the providers with which they contract, and so on. In the private employment sector, a change in jobs or employment status can of course result in a loss of coverage. But even within the comparatively stable contexts of Medicare Advantage and FEHBP, plan turnover is high, and provider participation also fluctuates substantially. All of this churning is costly, undermines continuity of care, and is difficult for enrollees, particularly those who require coordinated care.

The effects of plan turnover on the quality of care remain poorly understood, but important clues are provided by 2001 and 2002 surveys of people in Medicare whose private Medicare plans terminated or reduced service areas. The surveys found that health plan withdrawal not only harmed the finances of those affected, but also had negative effects on mental and physical health, with the consequences most pronounced for the most vulnerable patients. In the 2001 survey, for example, 22 percent of those seeing a specialist reported they had to stop seeing their specialist, and “fifteen percent said they did not get some prescribed medication since leaving their former plan. Disabled individuals, those in fair or poor health, and people of color reported the most trouble with access to care.”

No less important, the greater stability of enrollment and provider participation gives public insurance a greater potential to reap the rewards of investments in prevention and general health improvement that may have up-front costs. One of the costs of the fragmentation of health insurance is that health plans may not benefit from measures that improve the health or long-term health expenditures of enrollees. It is important to recognize that the potential benefits to health plans go beyond monetary savings to
include the value of better health and well-being that such measures may produce, insofar as enrollees recognize these broader benefits and reward health plans for them. As Randall Cebul and his colleagues explain, “In principle, insurers could capture some of this value in the form of higher premiums, if they could count on long-term relationships with employer groups or individual policy holders . . . . But insurance companies cannot count on such long-term relationships with many or most insured individuals.” Cebul and his colleagues cite a recent study of diabetes management at a private HMO: The study showed a positive social return for diabetes management, but the private return for the plan was negative in the first years and zero over the course of a decade—in part because the turnover of enrollees meant that the plan did not reap the potential benefits of long-term health improvements.

A public plan with a relatively stable enrollment base would be best poised to make long-term investments in patients’ health that deliver financial and social benefits down the road. This is yet another respect in which a new public plan, working with Medicare and private plans, could spearhead the testing and evaluation of potential delivery-system and payment reform; the collection, reporting, and use of ongoing performance data; and the streamlining of paperwork and administration in ways that would not be possible without a core role for public insurance for nonelderly Americans.

PUBLIC PLAN CHOICE AS A BENCHMARK

Public and private insurance have distinct strengths and distinct weaknesses. Private insurance is generally more dynamic and flexible than public insurance, but at the same time less stable and more administratively complex and costly. Public insurance is better at spreading risks broadly—given the extreme concentration of medical costs, private plans inevitably have incentives to “cherry-pick” healthier patients—but this advantage carries with it the potential cost of a lesser capacity to adapt rapidly to changing technology or the distinctive personal circumstances of individuals. Thus, a public-private hybrid can provide an important check on both the public and private sectors, ensuring flexibility and stability, market accountability and democratic accountability, inclusive social protection and private innovation—in short, a broadened range of good, meaningful choices.

Moreover, public and private plans can learn from each other as they exploit their strengths and remedy their weaknesses. Expanded coverage of prescription drugs by Medicare HMOs, for example, demonstrated the feasibility of drug coverage for the elderly and helped to increase political pressure for drug coverage for all people with Medicare. The development of performance measures for Medicare private plans provided a template for projects testing comparable measures under the public Medicare plan. Similarly, innovations in coordinating care for elders with chronic illness in private plans have provided a useful foundation for care-coordination demonstrations in the public Medicare plan.

Meanwhile, private insurers have emulated Medicare’s prospective payment system for hospitals, physicians and nursing homes, and many of the early techniques of utilization review were first developed by Medicare and later diffused to private insurers. Recently,
Aetna, WellPoint, and other larger insurers have moved to ban payments for care that results in serious errors—following the lead of Medicare’s effort to stop paying the cost of treating bed sores, falls, and other preventable injuries and infections. And no one doubts that Medicare has provided an important fallback for elderly and disabled Americans who have substantial health needs or whose private plans exit the market or switch benefits or providers.

The Need for a Level Playing Field

For the competitive and learning advantages of public-private competition to be realized, public and private plans must compete side by side on a level playing field. The purpose of this paper is not to outline the proper relationship between public and private insurance within a national insurance pool—that is the topic of a separate forthcoming policy brief. But, at a minimum, a level playing field requires that (1) both the public and private plans offer a good, comparable basic package of benefits; (2) private plans are regulated to ensure that they accept all comers, guarantee renewability of coverage, offer similar rates to all enrollees, and do not limit coverage for preexisting conditions; (3) adequate monitoring be in place to ensure that selective marketing or disenrollment does not occur; and (4) payments to the plans are risk adjusted so that plans that enroll a large number of the highest-cost patients that account for most national health spending are not disadvantaged.

Additional steps that may be needed to ensure fair competition—some potentially controversial—include (1) paying private plans a mix of prospective and retrospective payments to simultaneously encourage them to provide care efficiently and indemnify them against high-cost enrollees (this could include “clawing back” some amount from private plans that end with a highly favorable mix of enrollees as well); (2) a “hold harmless” principle that enrollees in the public plan would not have to pay higher premiums than enrollees in private plans if the higher costs of the public plan were due to the disproportionate enrollment of higher-cost patients; and (3) automatically enrolling those in the pool who do not choose a plan into the public plan as the default source of coverage. (Automatic enrollment in the public plan for those not choosing a plan would help the public plan to obtain a broad mix of risk, which may be difficult otherwise, because of the tendency for less healthy enrollees to enroll in the public plan. In addition, it would give those who did not choose a plan the broadest potential selection of providers.)

On the other hand, private plans will need assurances that the public plan will not trim provider rates so much that the plans are subject to cost shifting that drives up their premiums (although, as noted, the extent of cost shifting is often exaggerated) and both the public and private plans will have to be governed by an administrative structure that ensures fair representation of the interests of contracting private plans as well as patients, providers, employers, and the public at large. In particular, competition between the public and private plans probably should not be governed by the agency running the public plan itself, but instead by some higher-level body, such as the new national health board envisioned in a number of leading proposals for reform.
Why Regulation of Private Insurance Is Not Enough

As this discussion suggests, some of the shortcomings of private plans, such as their strong incentive to select healthier enrollees, can be partially addressed through regulations and new payment policies, including the twin requirements of open enrollment and community rating and measures to risk-adjust payments to plans so their incentive to attract less costly patients is reduced. Nonetheless, even with such regulations and payment reforms, a public plan competing with private plans is essential.

First, as already discussed, the Medicare program outperforms private insurance on costs and access even when compared with private plans that are regulated to ensure broad coverage, such as plans in FEHBP and Medicare Advantage. Medicare Advantage not only exercises substantial regulatory authority over private plans, but also has invested increasing resources in risk adjustment. Yet while Medicare Advantage plans have delivered broader services and diversity of plan offerings, they have certainly not delivered lower costs. Instead, they have resulted in the federal government spending more on Medicare than it would have otherwise—excess costs are projected to total nearly $150 billion between 2009 and 2017. Although these costs are principally the fault of a flawed method for paying private plans (one that should not be emulated in a new national pool), they suggest that simply regulating or adjusting payments to private plans to reduce risk selection will not guarantee that private plans focus on value.

Second, as emphasized throughout this brief, a competing public plan is needed to set a benchmark for private plans even in the context of private insurance regulations and risk adjustment—neither of which can be expected to fully change the incentives of private plans. Just as our nation’s Founders wanted ambition to check ambition to ensure that the “parchment barriers” of the Constitution were adhered to, public plan choice is a source of “checks and balances” designed to ensure that private plans have to uphold high standards of performance. . . .
§ 7. Economic Theory Applied: Liability and Other Doctrines

§ 7.1. Section 1983 and Qualified Immunity


MR. JUSTICE WHITE delivered the opinion of the Court.

Respondent Navarette, an inmate of Soledad Prison in California . . . , [claimed that various prison officials] had failed to mail various items of correspondence [in 1971 and 1972] . . . . These items . . . included letters to legal assistance groups, law students, the news media, and inmates in other state prisons, as well as personal friends. Some of these items had been returned to Navarette, some the defendants had refused to send by registered mail as Navarette had requested, and, it was alleged, none of the items had ever reached the intended recipient. This “interference” or “confiscation” was asserted to have occurred because the . . . officers had “negligently and inadvertently” misapplied the prison mail regulations and because the supervisory officers had “negligent[ly]” failed to provide sufficient training and direction to their subordinates, all assertedly in violation of Navarette’s constitutional rights. . . .

. . . [P]etitioners argued that . . . they were immune from liability for damages under § 1983 and hence were entitled to judgment as a matter of law. The claim was not that they shared the absolute immunity accorded judges and prosecutors but that they were entitled to . . . qualified immunity . . . .

We agree with petitioners that as prison officials and officers, they were not absolutely immune from liability in this § 1983 damages suit and could rely only on . . . qualified immunity . . . .

We further held in Wood v. Strickland that “if the work of the schools is to go forward,” there must be a degree of immunity so that “public school officials understand that action taken in the good-faith fulfillment of their responsibilities and within the bounds of reason under all the circumstances will not be punished and that they need not exercise their discretion with undue timidity.” This degree of immunity would be unavailable, however, if the official “knew or reasonably should have known that the action he took within his sphere of official responsibility would violate the constitutional rights of the student affected . . . .”

. . . [T]he immunity defense would be unavailing to petitioners if the constitutional right allegedly infringed by them was clearly established at the time of their challenged conduct, if they knew or should have known of that right, and if they knew or should have known that their conduct violated the constitutional norm. Petitioners claim that in 1971 and 1972 when the conduct involved in this case took place there was no
established First Amendment right protecting the mailing privileges of state prisoners and that hence there was no such federal right about which they should have known. We are in essential agreement with petitioners in this respect and also agree that they were entitled to judgment as a matter of law.

In ruling that petitioners’ conduct had encroached on Navarette’s First Amendment rights, the Court of Appeals relied on two of its own decisions, one in 1973 and the other in 1974, as well as upon Martinez v. Procunier, a 1973 three-judge court opinion . . . . The court relied on no earlier opinions, and this Court, in affirming the judgment in Martinez v. Procunier, did so on the ground that the constitutional rights of the addressees of a prisoner’s correspondence were involved when prison officials interfered with a prisoner’s outgoing mail. The question of the rights of the prisoner himself was left open. . . . The Court referred to no relevant pronouncements by courts in the Ninth Circuit other than the one then under review; and it is apparent that Procunier, the defendant in the Martinez suit and in this one, was then maintaining that there was no established constitutional right protecting prison mail under which his mail regulations could be challenged.

[N]one of these cases [that respondent relies on] deals with the rights of convicted prisoners in their mail and none furnishes an adequate basis for claiming that in 1971 and 1972 there was a “clearly established” constitutional right protecting Navarette’s correspondence involved in this case. . . .


JUSTICE POWELL delivered the opinion of the Court.

II

. . . [O]ur decisions consistently have held that government officials are entitled to some form of immunity from suits for damages. As recognized at common law, public officers require this protection to shield them from undue interference with their duties and from potentially disabling threats of liability.

Our decisions have recognized immunity defenses of two kinds. For officials whose special functions or constitutional status requires complete protection from suit, we have recognized the defense of “absolute immunity.” The absolute immunity of legislators, in their legislative functions, and of judges, in their judicial functions, now is well settled. Our decisions also have extended absolute immunity to certain officials of the Executive Branch. These include prosecutors and similar officials, executive officers engaged in adjudicative functions, and the President of the United States.

For executive officials in general, however, our cases make plain that qualified immunity represents the norm. . . . [A] qualified immunity defense for high executives reflect[s] an attempt to balance competing values: not only the importance of a damages remedy to
protect the rights of citizens, but also “the need to protect officials who are required to exercise their discretion and the related public interest in encouraging the vigorous exercise of official authority.” Without discounting the adverse consequences of denying high officials an absolute immunity from private lawsuits alleging constitutional violations . . . we emphasized our expectation that insubstantial suits need not proceed to trial . . . .

IV

Even if they cannot establish that their official functions require absolute immunity, petitioners assert that public policy at least mandates an application of the qualified immunity standard that would permit the defeat of insubstantial claims without resort to trial. We agree.

A

The resolution of immunity questions inherently requires a balance between the evils inevitable in any available alternative. In situations of abuse of office, an action for damages may offer the only realistic avenue for vindication of constitutional guarantees. It is this recognition that has required the denial of absolute immunity to most public officers. At the same time, however, it cannot be disputed seriously that claims frequently run against the innocent as well as the guilty—at a cost not only to the defendant officials, but to society as a whole. These social costs include the expenses of litigation, the diversion of official energy from pressing public issues, and the deterrence of able citizens from acceptance of public office. Finally, there is the danger that fear of being sued will “dampen the ardor of all but the most resolute, or the most irresponsible [public officials], in the unflinching discharge of their duties.”

In identifying qualified immunity as the best attainable accommodation of competing values, . . . we relied on the assumption that this standard would permit “[i]nsubstantial lawsuits [to] be quickly terminated.” Yet petitioners advance persuasive arguments that the dismissal of insubstantial lawsuits without trial—a factor presupposed in the balance of competing interests struck by our prior cases—requires an adjustment of the “good faith” standard established by our decisions.

B

Qualified or “good faith” immunity is an affirmative defense that must be pleaded by a defendant official. Decisions of this Court have established that the “good faith” defense has both an “objective” and a “subjective” aspect. . . . [W]e have held that qualified immunity would be defeated if an official “knew or reasonably should have known that the action he took within his sphere of official responsibility would violate the constitutional rights of the [plaintiff], or if he took the action with the malicious intention to cause a deprivation of constitutional rights or other injury . . . .”

The subjective element of the good-faith defense frequently has proved incompatible with our admonition . . . that insubstantial claims should not proceed to trial. . . . [A]n
official’s subjective good faith has been considered to be a question of fact that some
courts have regarded as inherently requiring resolution by a jury.

. . . [B]are allegations of malice should not suffice to subject government officials either
to the costs of trial or to the burdens of broad-reaching discovery. We therefore hold that
government officials performing discretionary functions generally are shielded from
liability for civil damages insofar as their conduct does not violate clearly established
statutory or constitutional rights of which a reasonable person would have known.

Reliance on the objective reasonableness of an official’s conduct, as measured by
reference to clearly established law, should avoid excessive disruption of government and
permit the resolution of many insubstantial claims on summary judgment. On summary
judgment, the judge appropriately may determine, not only the currently applicable law,
but whether that law was clearly established at the time an action occurred. If the law at
that time was not clearly established, an official could not reasonably be expected to
anticipate subsequent legal developments, nor could he fairly be said to ‘know’ that the
law forbade conduct not previously identified as unlawful. Until this threshold immunity
question is resolved, discovery should not be allowed. If the law was clearly established,
the immunity defense ordinarily should fail, since a reasonably competent public official
should know the law governing his conduct. Nevertheless, if the official pleading the
defense claims extraordinary circumstances and can prove that he neither knew nor
should have known of the relevant legal standard, the defense should be sustained. But
again, the defense would turn primarily on objective factors.

By defining the limits of qualified immunity essentially in objective terms, we provide no
license to lawless conduct. The public interest in deterrence of unlawful conduct and in
compensation of victims remains protected by a test that focuses on the objective legal
reasonableness of an official’s acts. Where an official could be expected to know that
certain conduct would violate statutory or constitutional rights, he should be made to
hesitate; and a person who suffers injury caused by such conduct may have a cause of
action. But where an official’s duties legitimately require action in which clearly
established rights are not implicated, the public interest may be better served by action
taken “with independence and without fear of consequences.” . . .


JUSTICE O’CONNOR delivered the opinion of the Court.

In Lugar v. Edmondson Oil Co. (1982), we left open the question whether private
defendants charged with 42 U.S.C. § 1983 liability for invoking state replevin,
garnishment, and attachment statutes later declared unconstitutional are entitled to
qualified immunity from suit. We now hold that they are not.
I

This dispute arises out of a soured cattle partnership. In July 1986, respondent Bill Cole sought to dissolve his partnership with petitioner Howard Wyatt. When no agreement could be reached, Cole, with the assistance of an attorney, respondent John Robbins II, filed a state court complaint in replevin against Wyatt, accompanied by a replevin bond of $18,000.

At that time, Mississippi law provided that an individual could obtain a court order for seizure of property possessed by another by posting a bond and swearing to a state court that the applicant was entitled to that property and that the adversary “wrongfully took and detain[ed]” or wrongfully detain[ed]” the property. The statute gave the judge no discretion to deny a writ of replevin.

After Cole presented a complaint and bond, the court ordered the county sheriff to seize 24 head of cattle, a tractor, and certain other personal property from Wyatt. Several months later, after a postseizure hearing, the court dismissed Cole’s complaint in replevin and ordered the property returned to Wyatt. When Cole refused to comply, Wyatt brought suit . . . , challenging the constitutionality of the statute and seeking injunctive relief and damages from respondents, the county sheriff, and the deputies involved in the seizure.

II

. . . In Lugar, the Court considered the scope of § 1983 liability in the context of garnishment, prejudgment attachment, and replevin statutes. In that case, the Court held that private parties who attached a debtor’s assets pursuant to a state attachment statute were subject to § 1983 liability if the statute was constitutionally infirm. Noting that our garnishment, prejudgment attachment, and replevin cases established that private use of state laws to secure property could constitute “state action” . . . , the Court held that private defendants invoking a state-created attachment statute act “under color of state law” . . . if their actions are “fairly attributable to the State.” This requirement is satisfied . . . if two conditions are met. First, the “deprivation must be caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the State or by a person for whom the State is responsible.” Second, the private party must have “acted together with or . . . obtained significant aid from state officials” or engaged in conduct “otherwise chargeable to the State.” The Court found potential § 1983 liability . . . because the attachment scheme was created by the State and because the private defendants, in invoking the aid of state officials to attach the disputed property, were “willful participant[s] in joint activity with the State or its agents.” . . .

III

Section 1983 “creates a species of tort liability that on its face admits of no immunities.” Nonetheless, we have accorded certain government officials either absolute or qualified immunity from suit if the “tradition of immunity was so firmly rooted in the common law and was supported by such strong policy reasons that ‘Congress would have specifically
so provided had it wished to abolish the doctrine.”’’ If parties seeking immunity were shielded from tort liability when Congress enacted [§ 1983,] we infer from legislative silence that Congress did not intend to abrogate such immunities when it imposed liability for actions taken under color of state law. Additionally, irrespective of the common law support, we will not recognize an immunity available at common law if § 1983’s history or purpose counsel against applying it in § 1983 actions.

In determining whether there was an immunity at common law that Congress intended to incorporate . . . , we look to the most closely analogous torts—in this case, malicious prosecution and abuse of process. At common law, these torts provided causes of action against private defendants for unjustified harm arising out of the misuse of governmental processes.

Respondents do not contend that private parties who instituted attachment proceedings and who were subsequently sued for malicious prosecution or abuse of process were entitled to absolute immunity. And with good reason; although public prosecutors and judges were accorded absolute immunity at common law, such protection did not extend to complaining witnesses who, like respondents, set the wheels of government in motion by instigating a legal action.

Nonetheless, respondents argue that at common law, private defendants could defeat a malicious prosecution or abuse of process action if they acted without malice and with probable cause, and that we should therefore infer that Congress did not intend to abrogate such defenses . . . . We adopted similar reasoning in Pierson v. Ray (1967). There, we held that police officers sued for false arrest under § 1983 were entitled to the defense that they acted with probable cause and in good faith when making an arrest under a statute they reasonably believed was valid. We recognized this defense because peace officers were accorded protection from liability at common law if they arrested an individual in good faith, even if the innocence of such person were later established.

The rationale we adopted in Pierson is of no avail to respondents here. Even if there were sufficient common law support to conclude that respondents, like the police officers in Pierson, should be entitled to a good faith defense, that would still not entitle them to what they sought and obtained in the courts below: the qualified immunity from suit accorded government officials under Harlow.

In Harlow, we altered the standard of qualified immunity adopted in our prior § 1983 cases because we recognized that “[t]he subjective element of the good-faith defense frequently [had] prove[d] incompatible with our admonition . . . that insubstantial claims should not proceed to trial.” Because of the attendant harms to government effectiveness caused by lengthy judicial inquiry into subjective motivation, we concluded that “bare allegations of malice should not suffice to subject government officials either to the costs of trial or to the burdens of broad-reaching discovery.” Accordingly, we held that government officials performing discretionary functions are shielded from “liability for civil damages insofar as their conduct [did] not violate clearly established statutory or constitutional rights of which a reasonable person would have known.” This wholly
objective standard, we concluded, would “avoid excessive disruption of government and permit the resolution of many insubstantial claims on summary judgment.”

...Harlow “completely reformulated qualified immunity along principles not at all embodied in the common law”... Harlow established an “immunity from suit rather than a mere defense to liability,” which, like an absolute immunity, “is effectively lost if a case is erroneously permitted to go to trial.” Thus, ... the denial of qualified immunity [is] immediately appealable.

It is this type of objectively determined, immediately appealable immunity that respondents asserted below. But, as our precedents make clear, the reasons for recognizing such an immunity were based not simply on the existence of a good faith defense at common law, but on the special policy concerns involved in suing government officials. Reviewing these concerns, we conclude that the rationales mandating qualified immunity for public officials are not applicable to private parties.

Qualified immunity strikes a balance between compensating those who have been injured by official conduct and protecting government’s ability to perform its traditional functions. Accordingly, we have recognized qualified immunity for government officials where it was necessary to preserve their ability to serve the public good or to ensure that talented candidates were not deterred by the threat of damages suits from entering public service. In short, the qualified immunity recognized in Harlow acts to safeguard government, and thereby to protect the public at large, not to benefit its agents.

These rationales are not transferable to private parties. Although principles of equality and fairness may suggest, as respondents argue, that private citizens who rely unsuspectingly on state laws they did not create and may have no reason to believe are invalid should have some protection from liability, as do their government counterparts, such interests are not sufficiently similar to the traditional purposes of qualified immunity to justify such an expansion. Unlike school board members, or police officers, or Presidential aides, private parties hold no office requiring them to exercise discretion; nor are they principally concerned with enhancing the public good. Accordingly, extending Harlow qualified immunity to private parties would have no bearing on whether public

---

2 In arguing that respondents are entitled to qualified immunity under Harlow, the dissent mixes apples and oranges. Even if we were to agree with the dissent’s proposition that elements a plaintiff was required to prove as part of her case in chief could somehow be construed as a “‘defense,’” and that this “defense” entitles private citizens to some protection from liability, we cannot agree that respondents are entitled to immunity from suit under Harlow. One could reasonably infer from the fact that a plaintiff’s malicious prosecution or abuse of process action failed if she could not affirmatively establish both malice and want of probable cause that plaintiffs bringing an analogous suit under § 1983 should be required to make a similar showing to sustain a § 1983 cause of action. Alternatively, if one accepts the dissent’s characterization of the common law as establishing an affirmative “defense” for private defendants, then one could also conclude that private parties sued under § 1983 should likewise be entitled to assert an affirmative defense based on a similar showing of good faith and/or probable cause. In neither case, however, is it appropriate to make the dissent’s leap: that because these common law torts partially included an objective component—probable cause—private defendants sued under § 1983 should be entitled to the objectively determined, immediately appealable immunity from suit accorded certain government officials under Harlow.
officials are able to act forcefully and decisively in their jobs or on whether qualified applicants enter public service. Moreover, unlike with government officials performing discretionary functions, the public interest will not be unduly impaired if private individuals are required to proceed to trial to resolve their legal disputes. In short, the nexus between private parties and the historic purposes of qualified immunity is simply too attenuated to justify such an extension of our doctrine of immunity.

For these reasons, we can offer no relief today. . . . Qualified immunity . . . is [not] available for private defendants faced with § 1983 liability for invoking a state replevin, garnishment, or attachment statute. . . . In so holding, however, we do not foreclose the possibility that private defendants faced with § 1983 liability under Lugar could be entitled to an affirmative defense based on good faith and/or probable cause or that § 1983 suits against private, rather than governmental, parties could require plaintiffs to carry additional burdens. . . .

CHIEF JUSTICE REHNQUIST, with whom JUSTICE SOURTCR and JUSTICE THOMAS join, dissenting. *

The Court notes that we have recognized an immunity in the § 1983 context in two circumstances. The first is when a similarly situated defendant would have enjoyed an immunity at common law at the time § 1983 was adopted. The second is when important public policy concerns suggest the need for an immunity. Because I believe that both requirements, as explained in our prior decisions, are satisfied here, I dissent.

First, I think it is clear that at the time § 1983 was adopted, there generally was available to private parties a good-faith defense to the torts of malicious prosecution and abuse of process. 1 And while the Court is willing to assume as much, it thinks this insufficient to sustain respondents’ claim to an immunity because the “qualified immunity” respondents seek is not equivalent to such a “defense.”

But I think the Court errs in suggesting that the availability of a good-faith common-law defense at the time of § 1983’s adoption is not sufficient to support their claim to immunity. The case on which respondents principally rely, Pierson, considered whether a police officer sued under § 1983 for false arrest could rely on a showing of good faith in order to escape liability. And while this Court concluded that the officer could rely on his own good faith, based in large part on the fact that a good-faith defense had been available at common law, the Court was at best ambiguous as to whether it was recognizing a “defense” or an “immunity.” Any initial ambiguity, however, has certainly been eliminated by subsequent cases; there can be no doubt that it is a qualified immunity to which the officer is entitled. Similarly, in Wood v. Strickland (1975), we recognized that, “[a]lthough there have been differing emphases and formulations of the common-

* [For convenience, Rehnquist’s dissent has been moved up, ahead of Kennedy’s concurrence.]
1 Describing the common law as providing a “defense” is something of a misnomer—under the common law it was plaintiff’s burden to establish as elements of the tort both that the defendant acted with malice and without probable cause. Referring to the defendant as having a good-faith defense is a useful shorthand for capturing plaintiff’s burden and the related notion that a defendant could avoid liability by establishing either a lack of malice or the presence of probable cause.
Thus, unlike the Court, I think our prior precedent establishes that a demonstration that a good-faith defense was available at the time § 1983 was adopted does, in fact, provide substantial support for a contemporary defendant claiming that he is entitled to qualified immunity in the analogous § 1983 context. While we refuse to recognize a common-law immunity if § 1983’s history or purpose counsel against applying it, I see no such history or purpose that would so counsel here.

Indeed, I am at a loss to understand what is accomplished by today’s decision—other than a needlessly fastidious adherence to nomenclature—given that . . . a good-faith defense will be available . . . on remand. Respondents presumably will be required to show the traditional elements of a good-faith defense—either that they acted without malice or that they acted with probable cause. The first element . . . encompasses an inquiry into subjective intent . . . . But the second element . . . focuses principally on objective reasonableness. Thus, respondents can successfully defend this suit simply by establishing that their reliance on the replevin statute was objectively reasonable for someone with their knowledge of the circumstances. But this is precisely the showing that entitles a public official to immunity.2

Nor do I see any reason that this “defense” may not be asserted early in the proceedings on a motion for summary judgment, just as a claim to qualified immunity may be. Provided that the historical facts are not in dispute, the presence or absence of “probable cause” has long been acknowledged to be a question of law. And so I see no reason that the trial judge may not resolve a summary judgment motion premised on such a good-faith defense, just as we have encouraged trial judges to do with respect to qualified immunity claims. Thus, private defendants who have invoked a state attachment law are put in the same position whether we recognize that they are entitled to qualified immunity or if we instead recognize a good-faith defense. Perhaps the Court believes that the “defense” will be less amenable to summary disposition than will the “immunity”; perhaps it believes the defense . . . must be submitted to the jury. While I can see no reason why this would be so (given that probable cause is a legal question), if it is true, today’s decision will only . . . increase litigation costs needlessly for hapless defendants.

This, in turn, leads to the second basis on which we have previously recognized a qualified immunity—reasons of public policy. Assuming that some practical difference will result from recognizing a defense but not an immunity, I think such a step is neither dictated by our prior decisions nor desirable. It is true, as the Court points out, that in abandoning a strictly historical approach to § 1983 immunities we have often explained our decision to recognize an immunity in terms of the special needs of public officials. But those cases simply do not answer—because the question was not at issue—whether

---

2 . . . [R]elying on the subjective belief, rather than on an objective lack of probable cause, is clearly exceptional. I see no reason to base our decision whether to extend a contemporary, objectively based qualified immunity on the exceptional common-law case.
similar (or even completely unrelated) reasons of public policy would warrant immunity for private parties as well.

I believe there are such reasons. The normal presumption that attaches to any law is that society will be benefitted if private parties rely on that law to provide them a remedy, rather than turning to some form of private, and perhaps lawless, relief. In denying immunity to those who reasonably rely on presumptively valid state law, and thereby discouraging such reliance, the Court expresses confidence that today’s decision will not “unduly impair” the public interest. I do not share that confidence. I would have thought it beyond peradventure that there is strong public interest in encouraging private citizens to rely on valid state laws of which they have no reason to doubt the validity.

Second, as with the police officer making an arrest, I believe the private plaintiff’s lot is “not so unhappy” that he must forgo recovery of property he believes to be properly recoverable through available legal processes or to be “mulcted in damages,” if his belief turns out to be mistaken. For as one Court of Appeals has pointed out, it is at least passing strange to conclude that private individuals are acting “under color of law” because they invoke a state garnishment statute and the aid of state officers, but yet deny them the immunity to which those same state officers are entitled, simply because the private parties are not state employees. While some of the strangeness may be laid at the doorstep of our decision in Lugar, there is no reason to proceed still further down this path. Our § 1983 jurisprudence has gone very far afield indeed, when it subjects private parties to greater risk than their public counterparts, despite the fact that § 1983’s historic purpose was “to prevent state officials from using the cloak of their authority under state law to violate rights protected against state infringement.” . . .

JUSTICE KENNEDY, with whom JUSTICE SCALIA joins, concurring.

. . . I agree with what THE CHIEF JUSTICE writes in dissent respecting the historical origins of our qualified immunity jurisprudence but [believe that] the result reached by the Court is quite consistent [with that history].

Both the Court and the dissent recognize that our original decisions recognizing defenses and immunities to suits brought under 42 U.S.C. § 1983 rely on analogous limitations existing in the common law when § 1983 was enacted. . . . [I]n Pierson, we recognized that under § 1983 police officers sued for false arrest had available what we described as a “defense of good faith and probable cause,” based on their reasonable belief that the statute under which they acted was constitutional. Pierson allowed the defense because with respect to the analogous common-law tort, the Court decided that officers had available to them a similar defense. The good-faith and probable-cause defense evolved into our modern qualified-immunity doctrine.

Our immunity doctrine is rooted in historical analogy, based on the existence of common-law rules in 1871, rather than in “freewheeling policy choice[s].” In cases involving absolute immunity we adhere to that view, granting immunity to the extent consistent with historical practice. In the context of qualified immunity for public officials, however, we have diverged to a substantial degree from the historical standards. In
Harlow, we “completely reformulated qualified immunity along principles not at all embodied in the common law, replacing the inquiry into subjective malice so frequently required at common law with an objective inquiry into the legal reasonableness of the official action.” The transformation was justified by the special policy concerns arising from public officials’ exposure to repeated suits. The dissent in today’s case argues that similar considerations justify a transformation of common-law standards in the context of private-party defendants. With this I cannot agree.

We need not decide whether or not it was appropriate for the Court in Harlow to depart from history in the name of public policy, reshaping immunity doctrines in light of those policy considerations. But I would not extend that approach to other contexts. Harlow was decided at a time when the standards applicable to summary judgment made it difficult for a defendant to secure summary judgment regarding a factual question such as subjective intent, even when the plaintiff bore the burden of proof on the question; and in Harlow we relied on that fact in adopting an objective standard for qualified immunity. However, subsequent clarifications to summary-judgment law have alleviated that problem, by allowing summary judgment to be entered against a nonmoving party “who fails to make a showing sufficient to establish the existence of an element essential to that party’s case, and on which that party will bear the burden of proof at trial.” Under these principles . . . , the strength of factual allegations such as subjective bad faith can be tested at the summary-judgment stage.

It must be remembered that unlike the common-law judges whose doctrines we adopt, we are devising limitations to a remedial statute, enacted by the Congress, which “on its face does not provide for any immunities.” We have imported common-law doctrines in the past because of our conclusion that the Congress which enacted § 1983 acted in light of existing legal principles. That suggests, however, that we may not transform what existed at common law based on our notions of policy or efficiency.

My conclusions are a mere consequence of the historical principles described in the dissent of THE CHIEF JUSTICE. The common-law tort actions most analogous to the action commenced here were malicious prosecution and abuse of process. In both of the common-law actions, it was essential for the plaintiff to prove that the wrongdoer acted with malice and without probable cause. As THE CHIEF JUSTICE states, it is something of a misnomer to describe the common law as creating a good-faith defense; we are in fact concerned with the essence of the wrong itself, with the essential elements of the tort. The malice element required the plaintiff to show that the challenged action was undertaken with an unlawful purpose, though it did not require a showing of ill will towards the plaintiff. To establish the absence of probable cause, a plaintiff was required to prove that a reasonable person, knowing what the defendant did, would not have believed that the prosecution or suit was well grounded, or that the defendant had in fact acted with the belief that the suit or prosecution in question was without probable cause. Our cases on the subject, beginning with Harlow, diverge from the common law in two ways. First, as THE CHIEF JUSTICE acknowledges, modern qualified immunity does not turn upon the subjective belief of the defendant. Second, the immunity diverges from the common-law model by requiring the defendant, not the plaintiff, to bear the burden of proof on the probable-cause issue.
The decision to impose these requirements under a rule of immunity has implications, though, well beyond a mere determination that one party or the other is in a better position to bear the burden of proof. It implicates as well the law’s definition of the wrong itself. At common law the action lay because the essence of the wrong was an injury caused by a suit or prosecution commenced without probable cause or with knowledge that it was baseless. To cast the issue in terms of immunity, however, is to imply that a wrong was committed but that it cannot be redressed. The difference is fundamental, for at stake is the concept of what society considers proper conduct and what it does not. Beneath the nomenclature lie considerations of substance.

*Harlow* was cast as an immunity case, involving as it did suit against officers of the Government. And immunity, as distinct, say, from a defense on the merits or an element of the plaintiff’s cause of action, is a legal inquiry, decided by the court rather than a jury, and on which an interlocutory appeal is available to defendants. Whether or not it is correct to diverge in these respects from the common-law model when governmental agents are the defendants, we ought not to adopt an automatic rule that the same analysis applies in suits against private persons. By casting the rule as an immunity, we imply the underlying conduct was unlawful, a most debatable proposition in a case where a private citizen may have acted in good-faith reliance upon a statute. And as we have defined the immunity, we also eliminate from the case any demonstration of subjective good faith. Under the common law, however, if the plaintiff could prove subjective bad faith on the part of the defendant, he had gone far towards proving both malice and lack of probable cause. Moreover, the question of the defendant’s beliefs was almost always one for the jury.

It is true that good faith may be difficult to establish in the face of a showing that from an objective standpoint no reasonable person could have acted as the defendant did, and in many cases the result would be the same under either test. This is why . . . the instances where the probable cause turns on subjective intent [are] the exceptional case. . . . [But] in some cases eliminating the defense based on subjective good faith can make a real difference, and again the instant case of alleged reliance on a statute deemed valid provides the example. It seems problematic to say that a defendant should be relieved of liability under some automatic rule of immunity if objective reliance upon a statute is reasonable but the defendant in fact had knowledge of its invalidity. Because the burden of proof . . . is the plaintiff’s, the question may be resolved on summary judgment if the plaintiff cannot come forward with facts from which bad faith can be inferred. But the question is a factual one, and a plaintiff may rely on circumstantial rather than direct evidence to make his case. The rule, of course, also works in reverse, for the existence of a statute thought valid ought to allow a defendant to argue that he acted in subjective good faith and is entitled to exoneration no matter what the objective test is.

The distinction I draw is important because there is support in the common law for the proposition that a private individual’s reliance on a statute, prior to a judicial determination of unconstitutionality, is considered reasonable as a matter of law; and therefore under the circumstances of this case, lack of probable cause can only be shown through proof of subjective bad faith. Thus the subjective element dismissed as exceptional by the dissent may be the rule rather than the exception.
I join the opinion of the Court because I believe there is nothing contrary to what I say in that opinion. Though they described the issue before them as “good-faith immunity,” [the lower courts] treated the question as one of law. The Court of Appeals in particular placed heavy reliance on the policy considerations favoring a rule that citizens may rely on statutes presumed to be valid. The latter inquiry . . . , however, goes mainly to the question of objective reasonableness. I do not understand [the lower courts] to make an unequivocal finding that the respondents before us acted with subjective good faith . . . . Furthermore, the question on which we granted certiorari was the narrow one whether private defendants in § 1983 suits are entitled to the same qualified immunity applicable to public officials, which of course would be subject to the objective standard of Harlow. Under my view the answer to that question is no. Though it might later be determined that there is no triable issue of fact to save the plaintiff’s case in the matter now before us, on remand it ought to be open to him at least in theory to argue that the defendants’ bad faith eliminates any reliance on the statute, just as it ought to be open to the defendants to show good faith even if some construct of a reasonable person in the defendants’ position would have acted in a different way. . . .

§ 7.1.4. McKnight v. Rees, 88 F.3d 417 (6th Cir. 1996)

BOYCE F. MARTIN, JR., Circuit Judge.

Tennessee contracts with private corporations to operate its prison facilities. In their capacities as correctional officers, defendants Daryll Richardson and John Walker allegedly violated plaintiff Ronnie McKnight’s Eighth Amendment right to be free from cruel and unusual punishment. . . .

. . . McKnight alleged that these officers . . . subject[ed] him to tight restraints during his transport to another prison. He claims that the restraints caused him serious medical injury which actually required hospitalization. Finally, the complaint states that McKnight’s protestations were ignored, and that Richardson and Walker taunted him after he complained about the restraints.

Richardson and Walker moved to dismiss the complaint, asserting that they were entitled to qualified immunity pursuant to their job function as correctional officers. . . .

Section 1983 provides a cause of action against any person who, under color of state law, deprives an individual of any right, privilege, or immunity secured by the Constitution and federal law.3 . . . [The goal of deterrence is b]alanced against . . . the equally important need to ensure that the threat of personal liability does not inhibit public

3 Richardson and Walker do not dispute that they were acting under color of state law for the purposes of section 1983 liability, even though each are privately employed by Corrections Corporation of America. Under the test set forth by the Supreme Court in Lugar, it appears that these defendants can be liable under section 1983, and are acting “under color of state law.” Nonetheless, we simply assume without deciding this fact for the purposes of this appeal.
officials from vigorously performing their duties as representatives of the public interest. We are asked here whether accommodation of these competing interests warrants the extension of qualified immunity to corrections officers employed by a private corporation under contract with the State of Tennessee.

... [S]tate employed correctional officers do enjoy the protections of qualified immunity in carrying out their duties. ... [But t]he principles undergirding qualified immunity do not necessarily apply with equal force when a private actor seeks its invocation.

In Duncan v. Peck (6th Cir. 1988), this Court held that qualified immunity does not protect from subsequent liability private individuals who resort to state garnishment or prejudgment attachment procedures in pursuit of personal interests. ... The Supreme Court has embraced this position as well. 

When determining whether a private individual may claim qualified immunity as a defense to a section 1983 claim, ... Wyatt teaches that we must look first to determine whether the “tradition of immunity [is] ... firmly rooted in the common law.” However, ... Harlow ... “completely reformulated qualified immunity along principles not at all embodied in the common law.” Therefore, although we have already recognized that there is “no evidence that the common law ever extended [qualified] immunity to include private citizens,” we conclude that, when determining whether qualified immunity is appropriate, the traditions of common law immunity inform our analysis but the presence or absence of immunity found there is not necessarily dispositive.

In addition to examining the historical roots of immunity, we also determine whether strong public policy reasons support the recognition of qualified immunity in particular cases. ... [W]e must “examine the nature of the functions with which a particular official or class of officials has been lawfully entrusted” to determine whether we should recognize qualified immunity in a given case. Finally, when determining whether public policy supports the grant or denial of immunity, “[w]e do not have a license to establish immunities from [section] 1983 actions in the interests of what we judge to be sound public policy.” We turn now to a perusal of the various rationales utilized to support qualified immunity.

While discussing the rationale underlying the creation of immunity from suit in Forrester v. White, the Court [cautioned] that ... “the threat of liability can create perverse incentives that operate to inhibit officials in the proper performance of their duties. In many contexts, government officials are expected to make decisions that are impartial or imaginative, and that above all are informed by considerations other than the personal interests of the decisionmaker. ... When officials are threatened with personal liability for acts taken pursuant to their official duties, they may well be induced to act with an excess of caution or otherwise to skew their decisions in ways that result in less than full fidelity to the objective and independent criteria that ought to guide their conduct.”

Emphasizing that courts should adopt a “functional” approach to questions of immunity, the Forrester Court characterized the inquiry as one in which courts should “seek to
evaluate the effect that exposure to particular forms of liability would likely have on the appropriate exercise of [the relevant government] functions.”

In Wyatt, the Supreme Court held that private defendants may not invoke the protections of qualified immunity when charged with section 1983 liability after utilizing state replevin, garnishment, and attachment statutes later held to be unconstitutional. The Court explained that the rationales advanced in support of qualified immunity for government actors are “not transferable to private parties.” . . .

The broad language used by the Court in this passage might lead one to believe that qualified immunity simply is not available to private party defendants, regardless of the nature of the act for which they are sued. However, the Court expressly limited its holding to the “precise issue” of whether qualified immunity is available to private defendants faced with section 1983 liability after utilizing a state replevin, garnishment or attachment statute.

Accordingly, several circuits read Wyatt narrowly, as applying only to cases in which . . . private parties “invoked state law in pursuit of private ends,” which is precisely the situation this Court confronted in Duncan. However, where private defendants are fulfilling government contracts or following court orders, several circuits have held that the umbrella of qualified immunity protects the private actors. . . .

In Sherman (v. Four County Counseling Ctr. (7th Cir. 1993)), the plaintiff had been involuntarily detained and given anti-psychotic medication pursuant to court order at the Four County Counseling Center. After his release from the center, Sherman sued pursuant to section 1983 the officer who arrested him and applied for the psychiatric evaluation, the state judge who assigned him for evaluation, the attorney who represented him at his preliminary hearing, Cass County, and Four County Counseling Center. On appeal, the Seventh Circuit considered whether Four County was entitled to qualified immunity for its actions. Answering the question in the affirmative, the court noted that Four County had accepted and administered care to Sherman under the state judge’s emergency detention order, and had complied with procedures set forth by Indiana statutory law during Sherman’s involuntary detention. These facts, according to the court, warranted the extension of qualified immunity to Four County:

The policy justifications which underlie the doctrine of qualified immunity for government officials apply with full force to Four County’s activities here. Although it is a private hospital, Four County accepted and cared for a state mental patient committed on an emergency basis. . . . If the actions it took pursuant to court order subject it to suit, private hospitals might well refuse to accept involuntary patients. This refusal would increase the load on the strained resources of the state’s public hospitals. . . . One purpose of qualified immunity is to avoid discouraging public service. Denying qualified immunity to private hospitals in this situation would do just that.
The Seventh Circuit recently reaffirmed the *Sherman* holding in a case factually analogous to the case at bar. In *Williams v. O’Leary* (7th Cir. 1995), the court held that medical director and staff physician of the Stateville Correctional Center in Illinois were entitled to qualified immunity even though each was an employee of Correctional Medical Systems, a private company contracting with the state to provide medical services at state correctional facilities. . . . The court believed that the fact that the defendant was “performing duties that would otherwise have to be performed by a public official who would clearly have qualified immunity” weighed in favor of granting qualified immunity to the private defendant.

In *Eagon v. Elk City*, the Tenth Circuit recently reaffirmed its rule that “a private individual who performs a government function pursuant to a state order or request is entitled to qualified immunity if a state official would have been entitled to such immunity had he performed the function himself.” One of the defendants in *Eagon* was a private party in charge of granting or denying applications for Christmas displays in a city park in Elk City, Oklahoma pursuant to a city council resolution. The lawsuit arose out of the denial of an application for a Christmas display. . . . [T]he Tenth Circuit believed that qualified immunity protected the private defendant’s actions, concluding that the defendant “was performing a government function pursuant to a government request,” and therefore was entitled to qualified immunity for her actions. . . .

Broadly stated, the Supreme Court has “followed a ‘functional’ approach to immunity law.” Relying on this fact, some circuits emphasize the function performed by the private party in determining whether to extend qualified immunity to the private defendant. For example, the First Circuit has held that “private parties under contract to perform statutorily mandated governmental duties are entitled to qualified immunity as functionally government employees.” While we do not disagree that we must examine the function or nature of the conduct at issue as an initial matter, we do not believe the analysis begins and ends with asking the question whether the private party is performing a government function. We must also examine whether the “special policy concerns involved in suing government officials,” also support the grant of qualified immunity to private actors performing what are traditionally governmental functions, as is the case here. Therefore, we next discuss whether the public policy underpinnings of qualified immunity support extending its protections to the private defendants here . . . .

The public undoubtedly has an interest in maintaining secure and efficient correctional facilities. We need not recite the litany of public benefits traditionally associated with maintaining an effective penal system here. We pause only to note that contracting out that function may save money for the state’s coffers, in addition to the traditional safety, deterrence, retribution, and symbolic rationales trotted out at every round table argument over criminal punishment. . . . [C]orrectional officers working for a private, for-profit corporation that has contracted with the state are serving a public interest by operating Tennessee’s prison facilities.

Nonetheless, . . . while privately employed correctional officers are serving the public interest by maintaining a correctional facility, they are not principally motivated by a
desire to further the interests of the public at large. Rather, as employees of a private
corporation seeking to maximize profits, correctional officers act, at least in part, out of a
desire to maintain the profitability of the corporation for whom they labor, thereby
ensuring their own job security. While this arguably does not vitiate furtherance of the
public interest, it upsets the balance qualified immunity seeks to establish “between
compensating those who have been injured by official conduct and protecting
government’s ability to perform its traditional functions.” The balance struck by qualified
immunity, at least implicitly, contemplates a government actor acting for the good of the
state, not a private actor acting for the good of the pocketbook. With respect to cutting
corners on constitutional guarantees, one commentator has explained that
“[e]ntrepreneurial jailers benefit directly, in the form of increased profits, from every
dime not spent.”

We believe this increased threat of injury by violation of constitutional guarantees
counsels against granting qualified immunity to correctional officers employed by a
private corporation to run a state’s prison facilities. Because private corporations
operating correctional facilities are not “principally concerned with enhancing the public
good,” the appropriate balance between the vindication of constitutional guarantees and
the furtherance of the public interest cannot be struck by a liability rule which assumes as
its starting point a public employee acting primarily for the benefit of the public. To
maintain this balance, the liability rule must adjust, and we believe the rule falls
somewhere short of qualified immunity for these privately employed officers.

To be sure, our refusal to grant qualified immunity to private corrections officers will
result in increased litigation costs, which may ultimately be passed on to the state in form
of an increased contract price. However, the public interest in minimizing the costs of
maintaining efficient, workable prison systems is not necessarily at odds with this result.
This increased litigation cost may be offset, at least partially, by the reduced
administrative costs associated with monitoring the private correctional facilities to
ensure compliance with constitutional rights. To the extent that private enforcement
through section 1983 acts as an effective monitor of private correctional corporations,
states will realize reduced administrative monitoring costs because the threat of liability
(and therefore reduced profits) will force the private actor to internalize the monitoring
costs otherwise borne by the state. In short, the state may pay more to obtain private
correctional services, but less to monitor them on an ongoing basis.

Further, the denial of qualified immunity here will not deter talented candidates from
entering public service, another of the major concerns undergirding qualified immunity
doctrine. We make two comments on this issue. First, the simple fact is that correctional
officers, privately employed by a corporation contracting with the state, are not in
“public” service, although they are performing a service that undoubtedly benefits the

---

4 We do not seek by implication to question the actions or intent of these defendants. Rather, we are simply
pointing out that for-profit corporations—and indirectly the employees of those corporations—seek to
realize what the name implies, a profit. Accordingly, private corporations running correctional facilities
have a greater incentive to cut costs by infringing upon the constitutional rights of prisoners in order to
ensure the profitability of the enterprise.
public interest. The motivations for entering the field undeniably differ from those prompting, for example, a run for public office, which rarely is undertaken for financial reasons. Rather, the spur in the flank of private correctional corporations is, of course, the profitability of the enterprise. The denial of qualified immunity will not deter qualified firms from contracting with the state to manage correctional facilities. Instead, firms simply will take account of this fact when analyzing the corporate balance sheet, and contract with the state accordingly. As long as a profit can be realized from the venture, we have no fear that denying qualified immunity to these defendants will result in a dearth of qualified applicants seeking entry into the field.

In sum, we disagree somewhat with the approach taken by several other circuits, which simply asks whether the private actor is performing a traditional governmental function (often, as is the case here, because the state has decided to “privatize” that function by contracting it out), and if so, extends immunity’s protective cloak to cover the private actors. As we only recently explained, “the rationale for granting qualified immunity to public officials—that they [will] act decisively in their jobs for the public good without fear of being sued—[does] not apply with equal force to private parties acting for their own ends.” . . . [I]n the case before us, we believe the public policy underpinnings of qualified immunity—which strike a balance between compensation for constitutional injury and the public interest in effective government—do not apply with equal force to acts undertaken by employees of a private, for-profit corporation. We therefore refuse to grant qualified immunity to correctional officers employed by a private corporation to maintain prison facilities for the state.

It may be that the appropriate balance to be struck here is to permit the correctional officers to assert a good faith defense, rather than qualified immunity. However, that issue is not before this Court in this interlocutory appeal. We leave that issue for another day, hold only that correctional officers employed by private, for profit corporations in contract with the state are not entitled to qualified immunity. . . .

DAVID A. NELSON, Circuit Judge, concurring in the judgment.

If I believed that we were free to dispose of this appeal on public policy grounds, my vote would be [in favor of qualified immunity]. The notion that there is likely to be a meaningful difference between the behavior of a correctional officer whose paycheck comes from the State of Tennessee and the behavior of a correctional officer whose paycheck comes from the Corrections Corporation of America is a notion, it seems to me, that bespeaks a vaguely utopian view of the virtues of those who feed at the public trough. This view is not one I share. If it is sound public policy to offer qualified immunity from suit to the correctional officer employed by a public corporation, I think it is equally sound policy to offer qualified immunity to the officer employed by a private corporation, assuming both employees perform identical functions for identical reasons.

As a matter of circuit precedent, however, I believe that this panel is foreclosed from following the course to which my own view of public policy might point. In Duncan, as
the Supreme Court has recognized, “[t]he Sixth Circuit . . . rejected qualified immunity for private defendants sued under § 1983 but . . . established a good faith defense.” . . .

Except insofar as to the right to an interlocutory appeal is concerned, the courts’ increasingly benign attitude toward summary judgment proceedings may frequently mean that there will be little practical difference between the good faith defense that was recognized in Duncan and the qualified immunity “defense” (arguably a misnomer) recognized by our sister circuits. The good faith defense is not before us now, however, and I intimate no view as to whether the defendants can prevail on such a defense under the facts of this case. . . .

§ 7.1.5. Richardson v. McKnight, 521 U.S. 399 (1997)

JUSTICE BREYER delivered the opinion of the Court.

The issue before us is whether prison guards who are employees of a private prison management firm are entitled to a qualified immunity from suit by prisoners charging a violation of 42 U.S.C. § 1983. We hold that they are not. . . .

II

B

. . . Wyatt . . . did not answer the legal question before us, whether petitioners—two employees of a private prison management firm—enjoy a qualified immunity from suit under § 1983. It does tell us, however, to look both to history and to the purposes that underlie government employee immunity in order to find the answer.*

History does not reveal a “firmly rooted” tradition of immunity applicable to privately employed prison guards. Correctional services in the United States have undergone various transformations. Government-employed prison guards may have enjoyed a kind of immunity defense arising out of their status as public employees at common law. But correctional functions have never been exclusively public. Private individuals operated local jails in the 18th century, and private contractors were heavily involved in prison management during the 19th century.

During that time, some States, including southern States like Tennessee, leased their entire prison systems to private individuals or companies which frequently took complete control over prison management, including inmate labor and discipline. Private prison lease agreements (like inmate suits) seem to have been more prevalent after § 1983’s enactment, but we have found evidence that the common law provided mistreated prisoners in prison leasing States with remedies against mistreatment by those private lessors. See, e.g., Dade Coal Co. v. Haslett (Ga. 1889) (convict can recover from

---

* [This paragraph has been moved from the end of Part II.A]
contractor for injuries sustained while on lease to private company); *Boswell v. Barnhart* (Ga. 1895) (wife can recover from contractor for chain-gang-related death of husband); *Dalheim v. Lemon* (1891) (contractor liable for convict injuries); *Tillar v. Reynolds* (Ark. 1910) (work farm owner liable for inmate beating death); *Weigel v. Brown* (CA8 1912) (prison contractor liable for unlawful whipping); see also *Edwards v. Pocahontas* (1891) (inmate can recover from municipal corporation for injuries caused by poor jail conditions); *Hall v. O’Neil Turpentine Co.* (Fla. 1908) (private prison contractor and subcontractor liable to municipality for escaped prisoner under lease agreement). Yet, we have found no evidence that the law gave purely private companies or their employees any special immunity from such suits. The case on which the dissent rests its argument, *Williams v. Adams* (Mass. 1861) (which could not—without more—prove the existence of such a tradition and does not, moreover, clearly involve a private prison operator) actually supports our point. It suggests that no immunity from suit would exist for the type of intentional conduct at issue in this case. See *ibid.* (were “battery” at issue, the case would be of a different “character” and “the defendant might be responsible”); see *id.*, at 176 (making clear that case only involves claim of ordinary negligence for lack of heat and other items, not “gross negligence,” “implied malice,” or “intention to do the prisoner any bodily injury”).

Correctional functions in England have been more consistently public, but historical sources indicate that England relied upon private jailers to manage the detention of prisoners from the Middle Ages until well into the 18th century. The common law forbade those jailers to subject “their prisoners to any pain or torment,” whether through harsh confinement in leg irons, or otherwise. And it apparently authorized prisoner lawsuits to recover damages. Apparently the law did provide a kind of immunity for certain private defendants, such as doctors or lawyers who performed services at the behest of the sovereign. But we have found no indication of any more general immunity that might have applied to private individuals working for profit.

Our research, including the sources that the parties have cited, reveals that in the 19th century (and earlier) sometimes private contractors and sometimes government itself carried on prison management activities. And we have found no conclusive evidence of a historical tradition of immunity for private parties carrying out these functions. History therefore does not provide significant support for the immunity claim.

C

Whether the immunity doctrine’s purposes warrant immunity for private prison guards presents a closer question. *Wyatt*, consistent with earlier precedent, described the doctrine’s purposes as protecting “government’s ability to perform its traditional functions” by providing immunity where “necessary to preserve” the ability of government officials “to serve the public good or to ensure that talented candidates were not deterred by the threat of damages suits from entering public service.” Earlier precedent described immunity as protecting the public from unwarranted timidity on the part of public officials by, for example, “encouraging the vigorous exercise of official authority,” by contributing to “‘principled and fearless decision-making,’” and by responding to the concern that threatened liability would, in Judge Hand’s words,
“‘dampen the ardour of all but the most resolute, or the most irresponsible,’” public officials.

The guards argue that those purposes support immunity whether their employer is private or public. Since private prison guards perform the same work as state prison guards, they say, they must require immunity to a similar degree. To say this, however, is to misread this Court’s precedents. The Court has sometimes applied a functional approach in immunity cases, but only to decide which type of immunity—absolute or qualified—a public officer should receive. And it never has held that the mere performance of a governmental function could make the difference between unlimited § 1983 liability and qualified immunity, especially for a private person who performs a job without government supervision or direction. Indeed a purely functional approach bristles with difficulty, particularly since, in many areas, government and private industry may engage in fundamentally similar activities, ranging from electricity production, to waste disposal, to even mail delivery.

Petitioners’ argument also overlooks certain important differences that, from an immunity perspective, are critical. First, the most important special government immunity-producing concern—unwarranted timidity—is less likely present, or at least is not special, when a private company subject to competitive market pressures operates a prison. Competitive pressures mean not only that a firm whose guards are too aggressive will face damages that raise costs, thereby threatening its replacement, but also that a firm whose guards are too timid will face threats of replacement by other firms with records that demonstrate their ability to do both a safer and a more effective job.

These ordinary marketplace pressures are present here. The private prison guards before us work for a large, multistate private prison management firm. The firm is systematically organized to perform a major administrative task for profit. It performs that task independently, with relatively less ongoing direct state supervision. It must buy insurance sufficient to compensate victims of civil rights torts. And, since the firm’s first contract expires after three years, its performance is disciplined, not only by state review, but also by pressure from potentially competing firms who can try to take its place.

In other words, marketplace pressures provide the private firm with strong incentives to avoid overly timid, insufficiently vigorous, unduly fearful, or “nonarduous” employee job performance. And the contract’s provisions—including those that might permit employee indemnification and avoid many civil-service restrictions—grant this private firm freedom to respond to those market pressures through rewards and penalties that operate directly upon its employees. To this extent, the employees before us resemble those of other private firms and differ from government employees.

This is not to say that government employees, in their efforts to act within constitutional limits, will always, or often, sacrifice the otherwise effective performance of their duties. Rather, it is to say that government employees typically act within a different system. They work within a system that is responsible through elected officials to voters who, when they vote, rarely consider the performance of individual subdepartments or civil servants specifically and in detail. And that system is often characterized by
multidepartment civil service rules that, while providing employee security, may limit the incentives or the ability of individual departments or supervisors flexibly to reward, or to punish, individual employees. Hence a judicial determination that “effectiveness” concerns warrant special immunity-type protection in respect to this latter (governmental) system does not prove its need in respect to the former. Consequently, we can find no special immunity-related need to encourage vigorous performance.

Second, “privatization” helps to meet the immunity-related need “to ensure that talented candidates” are “not deterred by the threat of damages suits from entering public service.” It does so in part because of the comprehensive insurance-coverage requirements just mentioned. The insurance increases the likelihood of employee indemnification and to that extent reduces the employment-discouraging fear of unwarranted liability potential applicants face. Because privatization law also frees the private prison-management firm from many civil service law restraints, it permits the private firm, unlike a government department, to offset any increased employee liability risk with higher pay or extra benefits. In respect to this second government-immunity-related purpose then, it is difficult to find a special need for immunity, for the guards’ employer can operate like other private firms; it need not operate like a typical government department.

Third, lawsuits may well “‘distrac[t]’” these employees “‘from their . . . duties,’” but the risk of “distraction” alone cannot be sufficient grounds for an immunity. Our qualified immunity cases do not contemplate the complete elimination of lawsuit-based distractions. And it is significant that, here, Tennessee law reserves certain important discretionary tasks—those related to prison discipline, to parole, and to good time—for state officials. Given a continual and conceded need for deterring constitutional violations and our sense that the firm’s tasks are not enormously different in respect to their importance from various other publicly important tasks carried out by private firms, we are not persuaded that the threat of distracting workers from their duties is enough virtually by itself to justify providing an immunity. Moreover, Tennessee, which has itself decided not to extend sovereign immunity to private prison operators (and arguably appreciated that this decision would increase contract prices to some degree), can be understood to have anticipated a certain amount of distraction.

D

Our examination of history and purpose thus reveals nothing special enough about the job or about its organizational structure that would warrant providing these private prison guards with a governmental immunity.

III

We close with three caveats. First, we have focused only on questions of § 1983 immunity and have not addressed whether the defendants are liable under § 1983 even though they are employed by a private firm. Because the Court of Appeals assumed, but did not decide, § 1983 liability, it is for the District Court to determine whether, under this Court’s decision in Lugar, defendants actually acted “under color of state law.”
Second, we have answered the immunity question narrowly, in the context in which it arose. That context is one in which a private firm, systematically organized to assume a major lengthy administrative task (managing an institution) with limited direct supervision by the government, undertakes that task for profit and potentially in competition with other firms. The case does not involve a private individual briefly associated with a government body, serving as an adjunct to government in an essential governmental activity, or acting under close official supervision.

Third, Wyatt explicitly stated that it did not decide whether or not the private defendants before it might assert, not immunity, but a special “good-faith” defense. . . . Like the Court in Wyatt, and the Court of Appeals in this case, we do not express a view on this last-mentioned question. . . .

**JUSTICE SCALIA, with whom THE CHIEF JUSTICE, JUSTICE KENNEDY, and JUSTICE THOMAS join, dissenting.**

In *Navarette*, we held that state prison officials, including both supervisory and subordinate officers, are entitled to qualified immunity in a suit brought under 42 U.S.C. § 1983. Today the Court declares that this immunity is unavailable to employees of private prison management firms, who perform the same duties as state-employed correctional officials, who exercise the most palpable form of state police power, and who may be sued for acting “under color of state law.” This holding is supported neither by common-law tradition nor public policy, and contradicts our settled practice of determining § 1983 immunity on the basis of the public function being performed.

I

The doctrine of official immunity against damages actions under § 1983 is rooted in the assumption that that statute did not abolish those immunities traditionally available at common law. I agree with the Court, therefore, that we must look to history to resolve this case. I do not agree with the Court, however, that the petitioners’ claim to immunity is defeated if they cannot provide an actual case, antedating or contemporaneous with the enactment of § 1983, in which immunity was successfully asserted by a private prison guard. It is only the absence of such a case, and not any explicit rejection of immunity by any common-law court, that the Court relies upon. The opinion observes that private jailers existed in the 19th century, and that they were successfully sued by prisoners. But one could just as easily show that government-employed jailers were successfully sued at common law, often with no mention of possible immunity. Indeed, as far as my research has disclosed, there may be more case-law support for immunity in the private-jailer context than in the government-jailer context. The only pre-§ 1983 jailer-immunity case of any sort that I am aware of is *Williams v. Adams* (Mass. 1861), decided only 10 years before § 1983 became law. And that case, which explicitly acknowledged that the issue
of jailer immunity was “novel,” appears to have conferred immunity upon an independent contractor.\footnote{\textit{Williams} held that prisoners could not recover damages for negligence against the master of a house of correction. That official seems to have been no more a “public officer” than the head of a private company running a prison. For example, the governing statute provided that he was to be paid by the prisoners for his expenses in supporting and employing them, and in event of their default he was given an action indebitatus assumpsit for the sum due, “which shall be deemed to be his own proper debt.” If he failed to distribute to the prisoners those “rations or articles of food, soap, fuel, or other necessaries” directed by the county commissioner (or the mayor and aldermen of Boston), he was subject to a fine. The opinion in \textit{Williams} says that “[t]he master of the house of correction is not an independent public officer, having the same relations to those who are confined therein that a deputy sheriff has to the parties to a writ committed to him to serve.”}

The truth to tell, \textit{Navarette}, which established § 1983 immunity for state prison guards, did not trouble itself with history, as our later § 1983 immunity opinions have done, but simply set forth a policy prescription. At this stage in our jurisprudence it is irrational, and productive of harmful policy consequences, to rely upon lack of case support to create an artificial limitation upon the scope of a doctrine (prison-guard immunity) that was itself not based on case support. I say an artificial limitation, because the historical principles on which common-law immunity was based, and which are reflected in our jurisprudence, plainly cover the private prison guard if they cover the nonprivate. Those principles are two: (1) immunity is determined by function, not status, and (2) even more specifically, private status is not disqualifying.

“[O]ur cases clearly indicate that immunity analysis rests on functional categories, not on the status of the defendant.” Immunity “flows not from rank or title or ‘location within the Government,’ . . . but from the nature of the responsibilities of the individual official.” “Running through our cases, with fair consistency, is a ‘functional’ approach to immunity questions . . . . Under that approach, we examine the nature of the functions with which a particular official or class of officials has been lawfully entrusted, and we seek to evaluate the effect that exposure to particular forms of liability would likely have on the appropriate exercise of those functions.” The parties concede that petitioners perform a prototypically governmental function (enforcement of state-imposed deprivation of liberty), and one that gives rise to qualified immunity.

The point that function rather than status governs the immunity determination is demonstrated in a prison-guard case virtually contemporaneous with the enactment of § 1983. \textit{Alamango v. Board of Supervisors of Albany Cty.} (N.Y. 1881) held that supervisors charged under state law with maintaining a penitentiary were immune from prisoner lawsuits. Although they were not formally state officers, the court emphasized the irrelevance of this fact:

“The duty of punishing criminals is inherent in the Sovereign power. It may be committed to agencies selected for that purpose, but such agencies, while engaged
in that duty, stand so far in the place of the State and exercise its political authority, and do not act in any private capacity.”

Private individuals have regularly been accorded immunity when they perform a governmental function that qualifies. We have long recognized the absolute immunity of grand jurors, noting that like prosecutors and judges they must “exercise a discretionary judgment on the basis of evidence presented to them.” “It is the functional comparability of [grand jurors’] judgments to those of the judge that has resulted in [their] being referred to as ‘quasi-judicial’ officers, and their immunities being termed ‘quasi-judicial’ as well.” Likewise, witnesses who testify in court proceedings have enjoyed immunity, regardless of whether they were government employees. “[T]he common law,” we have observed, “provided absolute immunity from subsequent damages liability for all persons—governmental or otherwise—who were integral parts of the judicial process.” I think it highly unlikely that we would deny prosecutorial immunity to those private attorneys increasingly employed by various jurisdictions in this country to conduct high-visibility criminal prosecutions. There is no more reason for treating private prison guards differently.

II

Later in its opinion, the Court seeks to establish that there are policy reasons for denying to private prison guards the immunity accorded to public ones. As I have indicated above, I believe that history and not judicially analyzed policy governs this matter—but even on its own terms the Court’s attempted policy distinction is unconvincing. The Court suggests two differences between civil-service prison guards and those employed by private prison firms which preclude any “special” need to give the latter immunity. First, the Court says that “unwarranted timidity” on the part of private guards is less likely to be a concern, since their companies are subject to market pressures that encourage them to be effective in the performance of their duties. If a private firm does not maintain a proper level of order, the Court reasons, it will be replaced by another one—so there is no need for qualified immunity to facilitate the maintenance of order.

This is wrong for several reasons. First of all, it is fanciful to speak of the consequences of “market” pressures in a regime where public officials are the only purchaser, and other people’s money the medium of payment. Ultimately, one prison-management firm will be selected to replace another prison-management firm only if a decision is made by some political official not to renew the contract. This is a government decision, not a market choice. If state officers turn out to be more strict in reviewing the cost and performance of privately managed prisons than of publicly managed ones, it will only be

---

2 The Court cites Alamango for the proposition that there is “no cause of action against [a] private contractor where [the] contractor [is] designated [a] state instrumentality by statute.” The opinion in Alamango, however, does not cite any statutory designation of the supervisors as a “state instrumentality,” and does not rely on such a designation for its holding. It does identify the Board of Supervisors as “a mere instrumentality selected by the State,” but the same could be said of the prison management firm here (or the master of the house of corrections in Williams, see n.1). If one were to accept the Court’s distinguishing of this case, all that would be needed to change the outcome in the present suit is the pointless formality of designating the contractor a “state instrumentality”—hardly a rational resolution of the question before us.
because they have chosen to be so. The process can come to resemble a market choice only to the extent that political actors will such resemblance—that is, to the extent that political actors (1) are willing to pay attention to the issue of prison services, among the many issues vying for their attention, and (2) are willing to place considerations of cost and quality of service ahead of such political considerations as personal friendship, political alliances, in-state ownership of the contractor, etc. Secondly and more importantly, however, if one assumes a political regime that is bent on emulating the market in its purchase of prison services, it is almost certainly the case that, short of mismanagement so severe as to provoke a prison riot, price (not discipline) will be the predominating factor in such a regime’s selection of a contractor. A contractor’s price must depend upon its costs; lawsuits increase costs; and “fearless” maintenance of discipline increases lawsuits. The incentive to down-play discipline will exist, moreover, even in those States where the politicians’ zeal for market emulation and budget cutting has waned, and where prison-management contract renewal is virtually automatic: the more cautious the prison guards, the fewer the lawsuits, the higher the profits. In sum, it seems that “market-competitive” private prison managers have even greater need than civil-service prison managers for immunity as an incentive to discipline.

The Court’s second distinction between state and private prisons is that privatization “helps to meet the immunity-related need to ensure that talented candidates are not deterred by the threat of damages suits from entering public service” as prison guards. This is so because privatization brings with it (or at least has brought with it in the case before us) (1) a statutory requirement for insurance coverage against civil-rights claims, which assertedly “increases the likelihood of employee indemnification,” and (2) a liberation “from many civil service law restraints” which prevent increased employee risk from being “offset . . . with higher pay or extra benefits.” As for the former (civil-rights liability insurance): surely it is the availability of that protection, rather than its actual presence in the case at hand, which decreases (if it does decrease, which I doubt) the need for immunity protection. (Otherwise, the Court would have to say that a private prison-management firm that is not required to purchase insurance, and does not do so, is more entitled to immunity; and that a government-run prison system that does purchase insurance is less entitled to immunity.) And of course civil-rights liability insurance is no less available to public entities than to private employers. But the second factor—liberation from civil-service limitations—is the more interesting one. First of all, simply as a philosophical matter it is fascinating to learn that one of the prime justifications for § 1983 immunity should be a phenomenon (civil-service laws) that did not even exist when § 1983 was enacted and the immunity created. Also as a philosophical matter, it is poetic justice (or poetic revenge) that the Court should use one of the principal economic benefits of “prison out-sourcing”—namely, the avoidance of civil-service salary and tenure encrustations—as the justification for a legal rule rendering out-sourcing more expensive. Of course the savings attributable to out-sourcing will not be wholly lost as a

---

3 This is true even of successfully defended lawsuits, and even of lawsuits that have been insured against. The Court thinks it relevant to the factor I am currently discussing that the private prison-management firm “must buy insurance sufficient to compensate victims of civil rights torts.” Belief in the relevance of this factor must be traceable, ultimately, to belief in the existence of a free lunch. Obviously, as civil-rights claims increase, the cost of civil-rights insurance increases.
result of today’s holding; they will be transferred in part from the public to prisonerplaintiffs and to lawyers. It is a result that only the American Bar Association and the American Federation of Government Employees could love. But apart from philosophical fascination, this second factor is subject to the same objection as the first: governments need not have civil-service salary encrustations (or can exempt prisons from them); and hence governments, no more than private prison employers, have any need for § 1983 immunity.

There is one more possible rationale for denying immunity to private prison guards worth discussing, albeit briefly. It is a theory so implausible that the Court avoids mentioning it, even though it was the primary reason given in the Court of Appeals decision that the Court affirms. It is that officers of private prisons are more likely than officers of state prisons to violate prisoners’ constitutional rights because they work for a profit motive, and hence an added degree of deterrence is needed to keep these officers in line. The Court of Appeals offered no evidence to support its bald assertion that private prison guards operate with different incentives than state prison guards, and gave no hint as to how prison guards might possibly increase their employers’ profits by violating constitutional rights. One would think that private prison managers, whose § 1983 damages come out of their own pockets, as compared with public prison managers, whose § 1983 damages come out of the public purse, would, if anything, be more careful in training their employees to avoid constitutional infractions. And in fact, States having experimented with prison privatization commonly report that the overall caliber of the services provided to prisoners has actually improved in scope and quality.

*     *     *

In concluding, I must observe that since there is no apparent reason, neither in history nor in policy, for making immunity hinge upon the Court’s distinction between public and private guards, the precise nature of that distinction must also remain obscure. Is it privity of contract that separates the two categories—so that guards paid directly by the State are “public” prison guards and immune, but those paid by a prison-management company “private” prison guards and not immune? Or is it rather “employee” versus “independent contractor” status—so that even guards whose compensation is paid directly by the State are not immune if they are not also supervised by a state official? Or is perhaps state supervision alone (without direct payment) enough to confer immunity? Or is it (as the Court’s characterization of Alamango suggests) the formal designation of the guards, or perhaps of the guards’ employer, as a “state instrumentality” that makes the difference? Since, as I say, I see no sense in the public-private distinction, neither do I see what precisely it consists of.

Today’s decision says that two sets of prison guards who are indistinguishable in the ultimate source of their authority over prisoners, indistinguishable in the powers that they possess over prisoners, and indistinguishable in the duties that they owe toward prisoners, are to be treated quite differently in the matter of their financial liability. The only sure effect of today’s decision—and the only purpose, as far as I can tell—is that it will artificially raise the cost of privatizing prisons. Whether this will cause privatization to be prohibitively expensive, or instead simply divert state funds that could have been saved

Draft — do not distribute!
or spent on additional prison services, it is likely that taxpayers and prisoners will suffer as a consequence. Neither our precedent, nor the historical foundations of § 1983, nor the policies underlying § 1983, support this result.


Chief Justice REHNQUIST delivered the opinion of the Court.

We decide here whether the implied damages action first recognized in Bivens v. Six Unknown Fed. Narcotics Agents (1971) should be extended to allow recovery against a private corporation operating a halfway house under contract with the Bureau of Prisons. We decline to so extend Bivens.

Petitioner Correctional Services Corporation (CSC), under contract with the federal Bureau of Prisons (BOP), operates Community Corrections Centers and other facilities that house federal prisoners and detainees. Since the late 1980’s, CSC has operated Le Marquis Community Correctional Center (Le Marquis), a halfway house located in New York City. Respondent John E. Malesko is a former federal inmate who, having been convicted of federal securities fraud in December 1992, was sentenced to a term of 18 months’ imprisonment under the supervision of the BOP. During his imprisonment, respondent was diagnosed with a heart condition and treated with prescription medication. Respondent’s condition limited his ability to engage in physical activity, such as climbing stairs.

In February 1993, the BOP transferred respondent to Le Marquis where he was to serve the remainder of his sentence. Respondent was assigned to living quarters on the fifth floor. On or about March 1, 1994, CSC instituted a policy at Le Marquis requiring inmates residing below the sixth floor to use the staircase rather than the elevator to travel from the first-floor lobby to their rooms. There is no dispute that respondent was exempted from this policy on account of his heart condition. Respondent alleges that on March 28, 1994, however, Jorge Urena, an employee of CSC, forbade him to use the elevator to reach his fifth-floor bedroom. Respondent protested that he was specially permitted elevator access, but Urena was adamant. Respondent then climbed the stairs, suffered a heart attack, and fell, injuring his left ear.

Three years after this incident occurred, respondent filed a pro se action against CSC and unnamed CSC employees. Two years later, now acting with counsel, respondent filed an amended complaint which named Urena as 1 of the 10 John Doe defendants. The amended complaint alleged that CSC, Urena, and unnamed defendants were “negligent in failing to obtain requisite medication for [respondent’s] condition and were further negligent by refusing [respondent] the use of the elevator.” It further alleged that respondent injured his left ear and aggravated a pre-existing condition “[a]s a result of the negligence of the Defendants.” Respondent demanded judgment in the sum of $1 million.
in compensatory damages, $3 million in anticipated future damages, and punitive 
damages “for such sum as the Court and/or [j]ury may determine.”

The District Court treated the amended complaint as raising claims under Bivens, and 
dismissed respondent’s cause of action in its entirety. Relying on our decision in FDIC v. 
Meyer (1994), the District Court reasoned that “a Bivens action may only be maintained 
against an individual,” and thus was not available against CSC, a corporate entity. With 
respect to Urena and the unnamed individual defendants, the complaint was dismissed on 
statute of limitations grounds. . . .

In Bivens, we recognized for the first time an implied private action for damages against 
federal officers alleged to have violated a citizen’s constitutional rights. Respondent now 
asks that we extend this limited holding to confer a right of action for damages against 
private entities acting under color of federal law. He contends that the Court must 
recognize a federal remedy at law wherever there has been an alleged constitutional 
depression, no matter that the victim of the alleged deprivation might have alternative 
remedies elsewhere, and that the proposed remedy would not significantly deter the 
principal wrongdoer, an individual private employee. We have heretofore refused to 
imply new substantive liabilities under such circumstances, and we decline to do so here.

Our authority to imply a new constitutional tort, not expressly authorized by statute, is 
anchored in our general jurisdiction to decide all cases “arising under the Constitution, 
laws, or treaties of the United States.” 28 U.S.C. § 1331. We first exercised this authority 
in Bivens, where we held that a victim of a Fourth Amendment violation by federal 
oficers may bring suit for money damages against the officers in federal court. Bivens 
acknowledged that Congress had never provided for a private right of action against 
federal officers, and that “the Fourth Amendment does not in so many words provide for 
its enforcement by award of money damages for the consequences of its violation.” 
Nonetheless, relying largely on earlier decisions implying private damages actions into 
federal statutes, and finding “no special factors counseling hesitation in the absence of 
affirmative action by Congress,” we found an implied damages remedy available under 
the Fourth Amendment.

In the decade following Bivens, we recognized an implied damages remedy under the 
Due Process Clause of the Fifth Amendment, Davis v. Passman (1979), and the Cruel 
and Unusual Punishments Clause of the Eighth Amendment, Carlson v. Green (1980). In 
both Davis and Carlson, we applied the core holding of Bivens, recognizing in limited 
circumstances a claim for money damages against federal officers who abuse their 
constitutional authority. In Davis, we inferred a new right of action chiefly because the 
plaintiff lacked any other remedy for the alleged constitutional deprivation. In Carlson, 
we inferred a right of action against individual prison officials where the plaintiff’s only 
alternative was a Federal Tort Claims Act (FTCA) claim against the United States. We 
reasoned that the threat of suit against the United States was insufficient to deter the 
unconstitutional acts of individuals. We also found it “crystal clear” that Congress 
intended the FTCA and Bivens to serve as “parallel” and “complementary” sources of 
liability.
Since Carlson we have consistently refused to extend Bivens liability to any new context or new category of defendants. In Bush v. Lucas (1983), we declined to create a Bivens remedy against individual Government officials for a First Amendment violation arising in the context of federal employment. Although the plaintiff had no opportunity to fully remedy the constitutional violation, we held that administrative review mechanisms crafted by Congress provided meaningful redress and thereby foreclosed the need to fashion a new, judicially crafted cause of action. We further recognized Congress’ institutional competence in crafting appropriate relief for aggrieved federal employees as a “special factor counseling hesitation in the creation of a new remedy.” We have reached a similar result in the military context, even where the defendants were alleged to have been civilian personnel.

In Schweiker v. Chilicky (1988), we declined to infer a damages action against individual Government employees alleged to have violated due process in their handling of Social Security applications. We observed that our “decisions have responded cautiously to suggestions that Bivens remedies be extended into new contexts.” In light of these decisions, we noted that “[t]he absence of statutory relief for a constitutional violation . . . does not by any means necessarily imply that courts should award money damages against the officers responsible for the violation.” We therefore rejected the claim that a Bivens remedy should be implied simply for want of any other means for challenging a constitutional deprivation in federal court. It did not matter, for example, that “[t]he creation of a Bivens remedy would obviously offer the prospect of relief for injuries that must now go unredressed.” So long as the plaintiff had an avenue for some redress, bedrock principles of separation of powers foreclosed judicial imposition of a new substantive liability.

Most recently, in FDIC v. Meyer, we unanimously declined an invitation to extend Bivens to permit suit against a federal agency, even though the agency—because Congress had waived sovereign immunity—was otherwise amenable to suit. Our opinion emphasized that “the purpose of Bivens is to deter the officer,” not the agency. We reasoned that if given the choice, plaintiffs would sue a federal agency instead of an individual who could assert qualified immunity as an affirmative defense. To the extent aggrieved parties had less incentive to bring a damages claim against individuals, “the deterrent effects of the Bivens remedy would be lost.” Accordingly, to allow a Bivens claim against federal agencies “would mean the evisceration of the Bivens remedy, rather than its extension.” We noted further that “special factors” counseled hesitation in light of the “potentially enormous financial burden” that agency liability would entail.

From this discussion, it is clear that the claim urged by respondent is fundamentally different from anything recognized in Bivens or subsequent cases. In 30 years of Bivens jurisprudence we have extended its holding only twice, to provide an otherwise nonexistent cause of action against individual officers alleged to have acted unconstitutionally, or to provide a cause of action for a plaintiff who lacked any alternative remedy for harms caused by an individual officer’s unconstitutional conduct. Where such circumstances are not present, we have consistently rejected invitations to extend Bivens, often for reasons that foreclose its extension here.
The purpose of *Bivens* is to deter individual federal officers from committing constitutional violations. *Meyer* made clear that the threat of litigation and liability will adequately deter federal officers for *Bivens* purposes no matter that they may enjoy qualified immunity, are indemnified by the employing agency or entity, or are acting pursuant to an entity’s policy. *Meyer* also made clear that the threat of suit against an individual’s employer was not the kind of deterrence contemplated by *Bivens*. This case is, in every meaningful sense, the same. For if a corporate defendant is available for suit, claimants will focus their collection efforts on it, and not the individual directly responsible for the alleged injury. On the logic of *Meyer*, inferring a constitutional tort remedy against a private entity like CSC is therefore foreclosed.

Respondent claims that even under *Meyer*’s deterrence rationale, implying a suit against private corporations acting under color of federal law is still necessary to advance the core deterrence purpose of *Bivens*. He argues that because corporations respond to market pressures and make decisions without regard to constitutional obligations, requiring payment for the constitutional harms they commit is the best way to discourage future harms. That may be so, but it has no relevance to *Bivens*, which is concerned solely with deterring the unconstitutional acts of individual officers. If deterring the conduct of a policymaking entity was the purpose of *Bivens*, then *Meyer* would have implied a damages remedy against the Federal Deposit Insurance Corporation; it was after all an agency policy that led to Meyer’s constitutional deprivation. But *Bivens* from its inception has been based not on that premise, but on the deterrence of individual officers who commit unconstitutional acts.

There is no reason for us to consider extending *Bivens* beyond this core premise here. To begin with, no federal prisoners enjoy respondent’s contemplated remedy. If a federal prisoner in a BOP facility alleges a constitutional deprivation, he may bring a *Bivens* claim against the offending individual officer, subject to the defense of qualified immunity. The prisoner may not bring a *Bivens* claim against the officer’s employer, the United States, or the BOP. With respect to the alleged constitutional deprivation, his only remedy lies against the individual; a remedy *Meyer* found sufficient, and which respondent did not timely pursue. Whether it makes sense to impose asymmetrical liability costs on private prison facilities alone is a question for Congress, not us, to decide.

Nor are we confronted with a situation in which claimants in respondent’s shoes lack effective remedies. It was conceded at oral argument that alternative remedies are at least

---

5 *JUSTICE STEVENS* claims that our holding in favor of CSC portends “tragic consequence[s],” and “jeopardize[s] the constitutional rights of . . . tens of thousands of inmates.” He refers to examples of cases suggesting that private correctional providers routinely abuse and take advantage of inmates under their control. In all but one of these examples, however, the private facility in question housed state prisoners—prisoners who already enjoy a right of action against private correctional providers under 42 U.S.C. § 1983. If it is true that the imperatives for deterring the unconstitutional conduct of private correctional providers are so strong as to demand that we imply a new right of action directly from the Constitution, then abuses of authority should be less prevalent in state facilities, where Congress already provides for such liability. That the trend appears to be just the opposite is not surprising given the BOP’s oversight and monitoring of its private contract facilities, which *JUSTICE STEVENS* does not mention.
as great, and in many respects greater, than anything that could be had under *Bivens*. For example, federal prisoners in private facilities enjoy a parallel tort remedy that is unavailable to prisoners housed in Government facilities. This case demonstrates as much, since respondent’s complaint in the District Court arguably alleged no more than a quintessential claim of negligence. It maintained that named and unnamed defendants were “*negligent* in failing to obtain requisite medication . . . and were further negligent by refusing . . . use of the elevator.” It further maintained that respondent suffered injuries “[a]s a result of the *negligence* of the Defendants.” The District Court, however, construed the complaint as raising a *Bivens* claim, presumably under the Cruel and Unusual Punishments Clause of the Eighth Amendment. Respondent accepted this theory of liability, and he has never sought relief on any other ground. This is somewhat ironic, because the heightened “deliberate indifference” standard of Eighth Amendment liability, *Estelle v. Gamble* (1976), would make it considerably more difficult for respondent to prevail than on a theory of ordinary negligence.

This also makes respondent’s situation altogether different from *Bivens*, in which we found alternative state tort remedies to be “inconsistent or even hostile” to a remedy inferred from the Fourth Amendment. When a federal officer appears at the door and requests entry, one cannot always be expected to resist. Yet lack of resistance alone might foreclose a cause of action in trespass or privacy. Therefore, we reasoned in *Bivens* that other than an implied constitutional tort remedy, “there remain[ed] . . . but the alternative of resistance, which may amount to a crime.” Such logic does not apply to respondent, whose claim of negligence or deliberate indifference requires no resistance to official action, and whose lack of alternative tort remedies was due solely to strategic choice.\(^6\)

Inmates in respondent’s position also have full access to remedial mechanisms established by the BOP, including suits in federal court for injunctive relief and grievances filed through the BOP’s Administrative Remedy Program. This program provides yet another means through which allegedly unconstitutional actions and policies can be brought to the attention of the BOP and prevented from recurring. And unlike the *Bivens* remedy, which we have never considered a proper vehicle for altering an entity’s policy, injunctive relief has long been recognized as the proper means for preventing entities from acting unconstitutionally.

In sum, respondent is not a plaintiff in search of a remedy as in *Bivens* and *Davis*. Nor does he seek a cause of action against an individual officer, otherwise lacking, as in *Carlson*. Respondent instead seeks a marked extension of *Bivens*, to contexts that would not advance *Bivens*’ core purpose of deterring individual officers from engaging in unconstitutional wrongdoing. The caution toward extending *Bivens* remedies into any new context, a caution consistently and repeatedly recognized for three decades, forecloses such an extension here. . . .

---

\(^6\) Where the government has directed a contractor to do the very thing that is the subject of the claim, we have recognized this as a special circumstance where the contractor may assert a defense. *Boyle v. United Technologies Corp.* (1988). The record here would provide no basis for such a defense.
JUSTICE STEVENS, with whom JUSTICE SOUTER, JUSTICE GINSBURG, and JUSTICE BREYER join, dissenting.

In Bivens, the Court affirmatively answered . . . whether a violation of the Fourth Amendment “by a federal agent acting under color of his authority gives rise to a cause of action for damages consequent upon his unconstitutional conduct.” Nearly a decade later, in Carlson, we held that a violation of the Eighth Amendment by federal prison officials gave rise to a Bivens remedy despite the fact that the plaintiffs also had a remedy against the United States under the Federal Tort Claims Act (FTCA). We stated: “Bivens established that the victims of a constitutional violation by a federal agent have a right to recover damages against the official in federal court despite the absence of any statute conferring such a right.”

In subsequent cases, we have decided that a Bivens remedy is not available for every conceivable constitutional violation. We have never, however, qualified our holding that Eighth Amendment violations are actionable under Bivens. Nor have we ever suggested that a category of federal agents can commit Eighth Amendment violations with impunity.

The parties before us have assumed that respondent’s complaint has alleged a violation of the Eighth Amendment. The violation was committed by a federal agent—a private corporation employed by the Bureau of Prisons to perform functions that would otherwise be performed by individual employees of the Federal Government. Thus, the question presented by this case is whether the Court should create an exception to the straightforward application of Bivens and Carlson, not whether it should extend our cases beyond their “core premise.” This point is evident from the fact that prior to our recent decision in Meyer, the Courts of Appeals had consistently and correctly held that corporate agents performing federal functions, like human agents doing so, were proper defendants in Bivens actions.

Meyer, which concluded that federal agencies are not suable under Bivens, does not lead to the outcome reached by the Court today. In that case, we did not discuss private corporate agents, nor suggest that such agents should be viewed differently from human ones. Rather, in Meyer, we drew a distinction between “federal agents” and “an agency of the Federal Government.” Indeed, our repeated references to the Federal Deposit Insurance Corporation’s (FDIC) status as a “federal agency” emphasized the FDIC’s affinity to the federal sovereign. We expressed concern that damages sought directly from federal agencies, such as the FDIC, would “creat[e] a potentially enormous financial burden for the Federal Government.” And it must be kept in mind that Meyer involved the FDIC’s waiver of sovereign immunity, which, had the Court in Meyer recognized a cause of action, would have permitted the very sort of lawsuit that Bivens presumed impossible: “a direct action against the Government.”

---

4 Meyer also did not address the present situation because the Court understood the plaintiff’s “real complaint” in that case to be that the individual officers would be shielded by qualified immunity, a
Moreover, in Meyer, as in Bush and Schweiker, we were not dealing with a well-recognized cause of action. The cause of action alleged in Meyer was a violation of procedural due process, and as the Meyer Court noted, “a Bivens action alleging a violation of the Due Process Clause of the Fifth Amendment may be appropriate in some contexts, but not in others.” Not only is substantive liability assumed in the present case, but respondent’s Eighth Amendment claim falls in the heartland of substantive Bivens claims.5

Because Meyer does not dispose of this case, the Court claims that the rationales underlying Bivens—namely, lack of alternative remedies and deterrence—are not present in cases in which suit is brought against a private corporation serving as a federal agent. However, common sense, buttressed by all of the reasons that supported the holding in Bivens, leads to the conclusion that corporate agents should not be treated more favorably than human agents.

First, the Court argues that respondent enjoys alternative remedies against the corporate agent that distinguish this case from Bivens. In doing so, the Court characterizes Bivens and its progeny as cases in which plaintiffs lacked “any alternative remedy.” In Bivens, however, even though the plaintiff’s suit against the Federal Government under state tort law may have been barred by sovereign immunity, a suit against the officer himself under state tort law was theoretically possible. Moreover, as the Court recognized in Carlson, Bivens plaintiffs also have remedies available under the FTCA. Thus, the Court is incorrect to portray Bivens plaintiffs as lacking any other avenue of relief, and to imply as a result that respondent in this case had a substantially wider array of non-Bivens remedies at his disposal than do other Bivens plaintiffs.6 If alternative remedies provide a sufficient justification for closing the federal forum here, where the defendant is a private corporation, the claims against the individual defendants in Carlson, in light of the FTCA alternative, should have been rejected as well.7

It is ironic that the Court relies so heavily for its holding on this assumption that alternative effective remedies—primarily negligence actions in state court—are available to respondent. Like Justice Harlan, I think it “entirely proper that these injuries be

5 The Court incorrectly assumes that we are being asked “to imply a new constitutional tort.” The tort here is, however, well established; the only question is whether a remedy in damages is available against a limited class of tortfeasors.

6 The Court recognizes that the question whether a Bivens action would lie against the individual employees of a private corporation like Correctional Services Corporation (CSC) is not raised in the present case. Both CSC and respondent have assumed Bivens would apply to these individuals, and the United States as amicus maintains that such liability would be appropriate under Bivens. It does seem puzzling that Bivens liability would attach to the private individual employees of such corporations—subagents of the Federal Government—but not to the corporate agents themselves. However, the United States explicitly maintains this to be the case, and the reasoning of the Court’s opinion relies, at least in part, on the availability of a remedy against employees of private prisons.

7 Although the Court lightly references administrative remedies that might be available to CSC-housed inmates, these are by no means the sort of comprehensive administrative remedies previously contemplated by the Court in Bush and Schweiker.
compensable according to uniform rules of federal law, especially in light of the very large element of federal law which must in any event control the scope of official defenses to liability.” And aside from undermining uniformity, the Court’s reliance on state tort law will jeopardize the protection of the full scope of federal constitutional rights. State law might have comparable causes of action for tort claims like the Eighth Amendment violation alleged here, but other unconstitutional actions by prison employees, such as violations of the Equal Protection or Due Process Clauses, may find no parallel causes of action in state tort law. Even though respondent here may have been able to sue for some degree of relief under state law because his Eighth Amendment claim could have been pleaded as negligence, future plaintiffs with constitutional claims less like traditional torts will not necessarily be so situated.8

Second, the Court claims that the deterrence goals of Bivens would not be served by permitting liability here. It cannot be seriously maintained, however, that tort remedies against corporate employers have less deterrent value than actions against their employees. As the Court has previously noted, the “organizational structure” of private prisons “is one subject to the ordinary competitive pressures that normally help private firms adjust their behavior in response to the incentives that tort suits provide—pressures not necessarily present in government departments.” Thus, the private corporate entity at issue here is readily distinguishable from the federal agency in Meyer. Indeed, a tragic consequence of today’s decision is the clear incentive it gives to corporate managers of privately operated custodial institutions to adopt cost-saving policies that jeopardize the constitutional rights of the tens of thousands of inmates in their custody.9

The Court raises a concern with imposing “asymmetrical liability costs on private prison facilities,” and further claims that because federal prisoners in Government-run institutions can only sue officers, it would be unfair to permit federal prisoners in private institutions to sue an “officer’s employer.” Permitting liability in the present case, however, would produce symmetry: both private and public prisoners would be unable to sue the principal (i.e., the Government), but would be able to sue the primary federal agent (i.e., the Government official or the corporation). Indeed, it is the Court’s decision that creates asymmetry—between federal and state prisoners housed in private correctional facilities. Under 42 U.S.C. § 1983, a state prisoner may sue a private prison

8 The Court blames respondent, who filed his initial complaint pro se, for the lack of state remedies in this case; according to the Court, respondent’s failure to bring a negligence suit in state court was “due solely to strategic choice.” Such strategic behavior, generally speaking, is imaginable, but there is no basis in the case before us to charge respondent with acting strategically. Respondent filed his complaint in federal court because he believed himself to have been severely maltreated while in federal custody, and he had no legal counsel to advise him to do otherwise. Without the aid of counsel, respondent not only failed to file for state relief, but he also failed to name the particular prison guard who was responsible for his injuries, resulting in the eventual dismissal of the claims against the individual officers as time barred. Respondent may have been an unsophisticated plaintiff, or, at worst, not entirely diligent about determining the identity of the guards, but it can hardly be said that “strategic choice” was the driving force behind respondent’s litigation behavior.

9 As amici for respondent explain, private prisons are exempt from much of the oversight and public accountability faced by the Bureau of Prisons, a federal entity. Indeed, because a private prison corporation’s first loyalty is to its stockholders, rather than the public interest, it is no surprise that cost-cutting measures jeopardizing prisoners’ rights are more likely in private facilities than in public ones.
for deprivation of constitutional rights, see Lugar (permitting suit under § 1983 against private corporations exercising “state action”), yet the Court denies such a remedy to that prisoner’s federal counterpart. It is true that we have never expressly held that the contours of Bivens and § 1983 are identical. The Court, however, has recognized sound jurisprudential reasons for parallelism, as different standards for claims against state and federal actors “would be incongruous and confusing.” The value of such parallelism was in fact furthered by Meyer, since § 1983 would not have provided the plaintiff a remedy had he pressed a similar claim against a state agency.

It is apparent from the Court’s critical discussion of the thoughtful opinions of Justice Harlan and his contemporaries, and from its erroneous statement of the question presented by this case as whether Bivens “should be extended” to allow recovery against a private corporation employed as a federal agent, that the driving force behind the Court’s decision is a disagreement with the holding in Bivens itself. There are at least two reasons why it is improper for the Court to allow its decision in this case to be influenced by that predisposition. First, as is clear from the legislative materials cited in Carlson, Congress has effectively ratified the Bivens remedy; surely Congress has never sought to abolish it. Second, a rule that has been such a well-recognized part of our law for over 30 years should be accorded full respect by the Members of this Court, whether or not they would have endorsed that rule when it was first announced. For our primary duty is to apply and enforce settled law, not to revise that law to accord with our own notions of sound policy. . . .

§ 7.2. Government Contractor Immunity Under the Federal Tort Claims Act


JUSTICE SCALIA delivered the opinion of the Court.

This case requires us to decide when a contractor providing military equipment to the Federal Government can be held liable under state tort law for injury caused by a design defect.

I

. . . Boyle, a United States Marine helicopter copilot, was killed when the CH-53D helicopter in which he was flying crashed off the coast of Virginia Beach, Virginia, during a training exercise. . . . Boyle’s father, petitioner here, [sued] the Sikorsky Division of United Technologies Corporation (Sikorsky), which built the helicopter for the United States. [Petitioner alleged product defects under Virginia law.] . . .
Petitioner’s broadest contention is that, in the absence of legislation specifically immunizing Government contractors from liability for design defects, there is no basis for judicial recognition of such a defense. We disagree. In most fields of activity, to be sure, this Court has refused to find federal pre-emption of state law in the absence of either a clear statutory prescription, or a direct conflict between federal and state law. But we have held that a few areas, involving “uniquely federal interests” are so committed by the Constitution and laws of the United States to federal control that state law is pre-empted and replaced, where necessary, by federal law of a content prescribed (absent explicit statutory directive) by the courts—so-called “federal common law.” . . .

. . . [T]he civil liabilities arising out of the performance of federal procurement contracts [are such an area of uniquely federal interest]. . . .

Moreover, it is plain that the Federal Government’s interest in the procurement of equipment is implicated by suits such as the present one—even though the dispute is one between private parties. It is true that where “litigation is purely between private parties and does not touch the rights and duties of the United States,” federal law does not govern. Thus, for example, in Miree v. DeKalb County (1977), which involved the question whether certain private parties could sue as third-party beneficiaries to an agreement between a municipality and the Federal Aviation Administration, we found that state law was not displaced because “the operations of the United States in connection with FAA grants such as these . . . would [not] be burdened” by allowing state law to determine whether third-party beneficiaries could sue, and because “any federal interest in the outcome of the [dispute] before us [was] far too speculative, far too remote a possibility to justify the application of federal law to transactions essentially of local concern.”

But the same is not true here. The imposition of liability on Government contractors will directly affect the terms of Government contracts: either the contractor will decline to manufacture the design specified by the Government, or it will raise its price. Either way, the interests of the United States will be directly affected.

That the procurement of equipment by the United States is an area of uniquely federal interest does not, however, end the inquiry. That merely establishes a necessary, not a sufficient, condition for the displacement of state law. Displacement will occur only where, as we have variously described, a “significant conflict” exists between an identifiable “federal policy or interest and the [operation] of state law,” or the application

---

2 As this language shows, Justice Brennan’s dissent is simply incorrect to describe Miree and other cases as declining to apply federal law despite the assertion of interests “comparable” to those before us here.

3 We refer here to the displacement of state law, although it is possible to analyze it as the displacement of federal-law reference to state law for the rule of decision. Some of our cases appear to regard the area in which a uniquely federal interest exists as being entirely governed by federal law, with federal law deigning to “borrow” or “incorporate” state law except where a significant conflict with federal policy exists. We see nothing to be gained by expanding the theoretical scope of the federal pre-emption beyond its practical effect, and so adopt the more modest terminology. If the distinction between displacement of state law and displacement of federal law’s incorporation of state law ever makes a practical difference, it at least does not do so in the present case.
of state law would “frustrate specific objectives” of federal legislation. The conflict with federal policy need not be as sharp as that which must exist for ordinary pre-emption when Congress legislates “in a field which the States have traditionally occupied.” Or to put the point differently, the fact that the area in question is one of unique federal concern changes what would otherwise be a conflict that cannot produce pre-emption into one that can. But conflict there must be. In some cases, for example where the federal interest requires a uniform rule, the entire body of state law applicable to the area conflicts and is replaced by federal rules. In others, the conflict is more narrow, and only particular elements of state law are superseded.

In *Miree*, the suit was not seeking to impose upon the person contracting with the Government a duty contrary to the duty imposed by the Government contract. Rather, it was the contractual duty itself that the private plaintiff (as third-party beneficiary) sought to enforce. Between *Miree* and the present case, it is easy to conceive of an intermediate situation, in which the duty sought to be imposed on the contractor is not identical to one assumed under the contract, but is also not contrary to any assumed. If, for example, the United States contracts for the purchase and installation of an air conditioning-unit, specifying the cooling capacity but not the precise manner of construction, a state law imposing upon the manufacturer of such units a duty of care to include a certain safety feature would not be a duty identical to anything promised the Government, but neither would it be contrary. The contractor could comply with both its contractual obligations and the state-prescribed duty of care. No one suggests that state law would generally be pre-empted in this context.

The present case, however, is at the opposite extreme from *Miree*. Here the state-imposed duty of care that is the asserted basis of the contractor’s liability (specifically, the duty to equip helicopters with the sort of escape-hatch mechanism petitioner claims was necessary) is precisely contrary to the duty imposed by the Government contract (the duty to manufacture and deliver helicopters with the sort of escape-hatch mechanism shown by the specifications). Even in this sort of situation, it would be unreasonable to say that there is always a “significant conflict” between the state law and a federal policy or interest. If, for example, a federal procurement officer orders, by model number, a quantity of stock helicopters that happen to be equipped with escape hatches opening outward, it is impossible to say that the Government has a significant interest in that particular feature. That would be scarcely more reasonable than saying that a private individual who orders such a craft by model number cannot sue for the manufacturer’s negligence because he got precisely what he ordered.

In its search for the limiting principle to identify those situations in which a “significant conflict” with federal policy or interests does arise, the Court of Appeals, in the lead case upon which its opinion here relied, identified as the source of the conflict the *Feres* doctrine, under which the Federal Tort Claims Act (FTCA) does not cover injuries to Armed Services personnel in the course of military service. See *Feres v. United States* (1950). Military contractor liability would conflict with this doctrine, the Fourth Circuit reasoned, since the increased cost of the contractor’s tort liability would be added to the price of the contract, and “[s]uch pass-through costs would . . . defeat the purpose of the immunity for military accidents conferred upon the government itself.” Other courts
upholding the defense have embraced similar reasoning. We do not adopt this analysis because it seems to us that the Feres doctrine, in its application to the present problem, logically produces results that are in some respects too broad and in some respects too narrow. Too broad, because if the Government contractor defense is to prohibit suit against the manufacturer whenever Feres would prevent suit against the Government, then even injuries caused to military personnel by a helicopter purchased from stock (in our example above), or by any standard equipment purchased by the Government, would be covered. Since Feres prohibits all service-related tort claims against the Government, a contractor defense that rests upon it should prohibit all service-related tort claims against the manufacturer—making inexplicable the three limiting criteria for contractor immunity (which we will discuss presently) that the Court of Appeals adopted. On the other hand, reliance on Feres produces (or logically should produce) results that are in another respect too narrow. Since that doctrine covers only service-related injuries, and not injuries caused by the military to civilians, it could not be invoked to prevent, for example, a civilian’s suit against the manufacturer of fighter planes, based on a state tort theory, claiming harm from what is alleged to be needlessly high levels of noise produced by the jet engines. Yet we think that the character of the jet engines the Government orders for its fighter planes cannot be regulated by state tort law, no more in suits by civilians than in suits by members of the Armed Services.

There is, however, a statutory provision that demonstrates the potential for, and suggests the outlines of, “significant conflict” between federal interests and state law in the context of Government procurement. In the FTCA, Congress authorized damages to be recovered against the United States for harm caused by the negligent or wrongful conduct of Government employees, to the extent that a private person would be liable under the law of the place where the conduct occurred. 28 U.S.C. § 1346(b). It excepted from this consent to suit, however,

“[a]ny claim . . . based upon the exercise or performance or the failure to exercise or perform a discretionary function or duty on the part of a federal agency or an employee of the Government, whether or not the discretion involved be abused.” 28 U.S.C. § 2680(a).

We think that the selection of the appropriate design for military equipment to be used by our Armed Forces is assuredly a discretionary function within the meaning of this provision. It often involves not merely engineering analysis but judgment as to the balancing of many technical, military, and even social considerations, including specifically the trade-off between greater safety and greater combat effectiveness. And we are further of the view that permitting “second-guessing” of these judgments through state tort suits against contractors would produce the same effect sought to be avoided by the FTCA exemption. The financial burden of judgments against the contractors would ultimately be passed through, substantially if not totally, to the United States itself, since defense contractors will predictably raise their prices to cover, or to insure against, contingent liability for the Government-ordered designs. To put the point differently: It makes little sense to insulate the Government against financial liability for the judgment that a particular feature of military equipment is necessary when the Government produces the equipment itself, but not when it contracts for the production. In sum, we
are of the view that state law which holds Government contractors liable for design defects in military equipment does in some circumstances present a “significant conflict” with federal policy and must be displaced.\(^5\)

We agree with the scope of displacement adopted by the Fourth Circuit here, which is also that adopted by the Ninth Circuit. Liability for design defects in military equipment cannot be imposed, pursuant to state law, when (1) the United States approved reasonably precise specifications; (2) the equipment conformed to those specifications; and (3) the supplier warned the United States about the dangers in the use of the equipment that were known to the supplier but not to the United States. The first two of these conditions assure that the suit is within the area where the policy of the “discretionary function” would be frustrated—i.e., they assure that the design feature in question was considered by a Government officer, and not merely by the contractor itself. The third condition is necessary because, in its absence, the displacement of state tort law would create some incentive for the manufacturer to withhold knowledge of risks, since conveying that knowledge might disrupt the contract but withholding it would produce no liability. We adopt this provision lest our effort to protect discretionary functions perversely impede them by cutting off information highly relevant to the discretionary decision.

We have considered the alternative formulation of the Government contractor defense, urged upon us by petitioner, which was adopted by the Eleventh Circuit . . . . That would preclude suit only if (1) the contractor did not participate, or participated only minimally, in the design of the defective equipment; or (2) the contractor timely warned the Government of the risks of the design and notified it of alternative designs reasonably known by it, and the Government, although forewarned, clearly authorized the contractor to proceed with the dangerous design. While this formulation may represent a perfectly reasonable tort rule, it is not a rule designed to protect the federal interest embodied in the “discretionary function” exemption. The design ultimately selected may well reflect a significant policy judgment by Government officials whether or not the contractor rather than those officials developed the design. In addition, it does not seem to us sound policy to penalize, and thus deter, active contractor participation in the design process, placing the contractor at risk unless it identifies all design defects. . . .

**JUSTICE BRENNAN, with whom JUSTICE MARSHALL and JUSTICE BLACKMUN join, dissenting.**

Lieutenant David A. Boyle died when the CH-53D helicopter he was copiloting spun out of control and plunged into the ocean. We may assume, for purposes of this case, that Lt. Boyle was trapped under water and drowned because respondent United Technologies negligently designed the helicopter’s escape hatch. We may further assume that any competent engineer would have discovered and cured the defects, but that they inexplicably escaped respondent’s notice. Had respondent designed such a death trap for a commercial firm, Lt. Boyle’s family could sue under Virginia tort law and be

\(^5\) **JUSTICE BRENNAN**’s assumption that the outcome of this case would be different if it were brought under the Death on the High Seas Act (1920) is not necessarily correct. That issue is not before us, and we think it inappropriate to decide it in order to refute (or, for that matter, to construct) an alleged inconsistency.
compensated for his tragic and unnecessary death. But respondent designed the helicopter for the Federal Government, and that, the Court tells us today, makes all the difference: Respondent is immune from liability so long as it obtained approval of “reasonably precise specifications”—perhaps no more than a rubber stamp from a federal procurement officer who might or might not have noticed or cared about the defects, or even had the expertise to discover them.

If respondent’s immunity “bore the legitimacy of having been prescribed by the people’s elected representatives,” we would be duty bound to implement their will, whether or not we approved. Congress, however, has remained silent—and conspicuously so, having resisted a sustained campaign by Government contractors to legislate for them some defense. The Court—unelected and unaccountable to the people—has unabashedly stepped into the breach to legislate a rule denying Lt. Boyle’s family the compensation that state law assures them. This time the injustice is of this Court’s own making.

Worse yet, the injustice will extend far beyond the facts of this case, for the Court’s newly discovered Government contractor defense is breathtakingly sweeping. It applies not only to military equipment like the CH-53D helicopter, but (so far as I can tell) to any made-to-order gadget that the Federal Government might purchase after previewing plans—from NASA’s Challenger space shuttle to the Postal Service’s old mail cars. The contractor may invoke the defense in suits brought not only by military personnel like Lt. Boyle, or Government employees, but by anyone injured by a Government contractor’s negligent design, including, for example, the children who might have died had respondent’s helicopter crashed on the beach. It applies even if the Government has not intentionally sacrificed safety for other interests like speed or efficiency, and, indeed, even if the equipment is not of a type that is typically considered dangerous; thus, the contractor who designs a Government building can invoke the defense when the elevator cable snaps or the walls collapse. And the defense is invocable regardless of how blatant or easily remedied the defect, so long as the contractor missed it and the specifications approved by the Government, however unreasonably dangerous, were “reasonably precise.”

In my view, this Court lacks both authority and expertise to fashion such a rule, whether to protect the Treasury of the United States or the coffers of industry. Because I would leave that exercise of legislative power to Congress, where our Constitution places it, I would reverse the Court of Appeals and reinstate petitioner’s jury award.

II

Congress has not decided to supersede state law here (if anything, it has decided not to) and the Court does not pretend that its newly manufactured “Government contractor defense” fits within any of the handful of “narrow areas” of “uniquely federal interests” in which we have heretofore done so. Rather, the Court creates a new category of “uniquely federal interests” out of a synthesis of two whose origins predate Erie itself: the interest in administering the “obligations to and rights of the United States under its contracts,” and the interest in regulating the “civil liability of federal officials for actions taken in the course of their duty.” This case is, however, simply a suit between two
private parties. We have steadfastly declined to impose federal contract law on relationships that are collateral to a federal contract, or to extend the federal employee’s immunity beyond federal employees. And the Court’s ability to list 2, or 10, inapplicable areas of “uniquely federal interest” does not support its conclusion that the liability of Government contractors is so “clear and substantial” an interest that this Court must step in lest state law does “major damage.”

A

The proposition that federal common law continues to govern the “obligations to and rights of the United States under its contracts” is nearly as old as *Erie* itself. Federal law typically controls when the Federal Government is a party to a suit involving its rights or obligations under a contract, whether the contract entails procurement, a loan, a conveyance of property, or a commercial instrument issued by the Government or assigned to it. Any such transaction necessarily “radiate[s] interests in transactions between private parties.” But it is by now established that our power to create federal common law controlling the Federal Government’s contractual rights and obligations does not translate into a power to prescribe rules that cover all transactions or contractual relationships collateral to Government contracts.

In *Miree*, for example, the county was contractually obligated under a grant agreement with the Federal Aviation Administration (FAA) to “restrict the use of land adjacent to . . . the Airport to activities and purposes compatible with normal airport operations including landing and takeoff of aircraft.” At issue was whether the county breached its contractual obligation by operating a garbage dump adjacent to the airport, which allegedly attracted the swarm of birds that caused a plane crash. Federal common law would undoubtedly have controlled in any suit by the Federal Government to enforce the provision against the county or to collect damages for its violation. The diversity suit, however, was brought not by the Government, but by assorted private parties injured in some way by the accident. We observed that “the operations of the United States in connection with FAA grants such as these are undoubtedly of considerable magnitude,” and that “the United States has a substantial interest in regulating aircraft travel and promoting air travel safety.” Nevertheless, we held that state law should govern the claim because “only the rights of private litigants are at issue here,” and the claim against the county “will have no direct effect upon the United States or its Treasury” (emphasis added).

*Miree* relied heavily on *Parnell* and *Wallis v. Pan American Petroleum Corp.*, the former involving commercial paper issued by the United States and the latter involving property rights in federal land. In the former case, *Parnell* cashed certain bonds guaranteed by the Government that had been stolen from their owner, a bank. It is beyond dispute that federal law would have governed the United States’ duty to pay the value bonds upon presentation; we held as much in *Clearfield Trust*. But the central issue in *Parnell*, a diversity suit, was whether the victim of the theft could recover the money paid to Parnell. That issue, we held, was governed by state law, because the “litigation [was] purely between private parties and [did] not touch the rights and duties of the United States.”
The same was true in *Wallis*, which also involved a Government contract—a lease issued by the United States to a private party under the Mineral Leasing Act of 1920—governed entirely by federal law. Again, the relationship at issue in this diversity case was collateral to the Government contract: It involved the validity of contractual arrangements between the lessee and other private parties, not between the lessee and the Federal Government. Even though a federal statute authorized certain assignments of lease rights and imposed certain conditions on their validity, we held that state law, not federal common law, governed their validity because application of state law would present “no significant threat to any identifiable federal policy or interest.”

Here, as in *Miree*, *Parnell*, and *Wallis*, a Government contract governed by federal common law looms in the background. But here, too, the United States is not a party to the suit and the suit neither “touch(es) the rights and duties of the United States,” nor has a “direct effect upon the United States or its Treasury.” The relationship at issue is at best collateral to the Government contract. We have no greater power to displace state law governing the collateral relationship in the Government procurement realm than we had to dictate federal rules governing equally collateral relationships in the areas of aviation, Government-issued commercial paper, or federal lands.

That the Government might have to pay higher prices for what it orders if delivery in accordance with the contract exposes the seller to potential liability does not distinguish this case. Each of the cases just discussed declined to extend the reach of federal common law despite the assertion of comparable interests that would have affected the terms of the Government contract—whether its price or its substance—just as “directly” (or indirectly). Third-party beneficiaries can sue under a county’s contract with the FAA, for example, even though—as the Court’s focus on the absence of “direct effect on the United States or its Treasury” suggests—counties will likely pass on the costs to the Government in future contract negotiations. Similarly, we held that state law may govern the circumstances under which stolen federal bonds can be recovered, notwithstanding *Parnell*’s argument that “the value of bonds to the first purchaser and hence their salability by the Government would be materially affected.” As in each of the cases declining to extend the traditional reach of federal law of contracts beyond the rights and duties of the Federal Government, “any federal interest in the outcome of the question before us ‘is far too speculative, far too remote a possibility to justify the application of federal law to transactions essentially of local concern.’”

B

Our “uniquely federal interest” in the tort liability of affiliates of the Federal Government is equally narrow. The immunity we have recognized has extended no further than a subset of “officials of the Federal Government” and has covered only “discretionary” functions within the scope of their legal authority. Never before have we so much as intimated that the immunity (or the “uniquely federal interest” that justifies it) might extend beyond that narrow class to cover also nongovernment employees whose authority to act is independent of any source of federal law and that are as far removed from the “functioning of the Federal Government” as is a Government contractor.
The historical narrowness of the federal interest and the immunity is hardly accidental. A federal officer exercises statutory authority, which not only provides the necessary basis for the immunity in positive law, but also permits us confidently to presume that interference with the exercise of discretion undermines congressional will. In contrast, a Government contractor acts independently of any congressional enactment. Thus, immunity for a contractor lacks both the positive law basis and the presumption that it furthers congressional will.

Moreover, even within the category of congressionally authorized tasks, we have deliberately restricted the scope of immunity to circumstances in which “the contributions of immunity to effective government in particular contexts outweigh the perhaps recurring harm to individual citizens,” because immunity “contravenes the basic tenet that individuals be held accountable for their wrongful conduct.” The extension of immunity to Government contractors skews the balance we have historically struck. On the one hand, whatever marginal effect contractor immunity might have on the “effective administration of policies of government,” its “harm to individual citizens” is more severe than in the Government-employee context. Our observation that “there are . . . other sanctions than civil tort suits available to deter the executive official who may be prone to exercise his functions in an unworthy and irresponsible manner” offers little deterrence to the Government contractor. On the other hand, a grant of immunity to Government contractors could not advance “the fearless, vigorous, and effective administration of policies of government” nearly as much as does the current immunity for Government employees. In the first place, the threat of a tort suit is less likely to influence the conduct of an industrial giant than that of a lone civil servant, particularly since the work of a civil servant is significantly less profitable, and significantly more likely to be the subject of a vindictive lawsuit. In fact, were we to take seriously the Court’s assertion that contractors pass their costs—including presumably litigation costs—through, “substantially if not totally, to the United States,” the threat of a tort suit should have only marginal impact on the conduct of Government contractors. More importantly, inhibition of the Government official who actually sets Government policy presents a greater threat to the “administration of policies of government,” than does inhibition of a private contractor, whose role is devoted largely to assessing the technological feasibility and cost of satisfying the Government’s predetermined needs. Similarly, unlike tort suits against Government officials, tort suits against Government contractors would rarely “consume time and energies” that “would otherwise be devoted to governmental service.”

In short, because the essential justifications for official immunity do not support an extension to the Government contractor, it is no surprise that we have never extended it that far . . .

III

[T]he Court [also] invokes the discretionary function exception of the Federal Tort Claims Act (FTCA), 28 U.S.C. § 2680(a). The Court does not suggest that the exception has any direct bearing here, for petitioner has sued a private manufacturer (not the Federal Government) under Virginia law (not the FTCA). Perhaps that is why respondent
has three times disavowed any reliance on the discretionary function exception, even after coaching by the Court, as has the Government.

Notwithstanding these disclaimers, the Court invokes the exception, reasoning that federal common law must immunize Government contractors from state tort law to prevent erosion of the discretionary function exception’s policy of foreclosing judicial “‘second-guessing’” of discretionary governmental decisions. The erosion the Court fears apparently is rooted not in a concern that suits against Government contractors will prevent them from designing, or the Government from commissioning the design of, precisely the product the Government wants, but in the concern that such suits might preclude the Government from purchasing the desired product at the price it wants: “The financial burden of judgments against the contractors,” the Court fears, “would ultimately be passed through, substantially if not totally, to the United States itself.”

Even granting the Court’s factual premise, which is by no means self-evident, the Court cites no authority for the proposition that burdens imposed on Government contractors, but passed on to the Government, burden the Government in a way that justifies extension of its immunity. However substantial such indirect burdens may be, we have held in other contexts that they are legally irrelevant.

Moreover, the statutory basis on which the Court’s rule of federal common law totters is more unstable than any we have ever adopted. In the first place, we rejected an analytically similar attempt to construct federal common law out of the FTCA when we held that the Government’s waiver of sovereign immunity for the torts of its employees does not give the Government an implied right of indemnity from them, even though the “[t]he financial burden placed on the United States by the Tort Claims Act [could conceivably be] so great that government employees should be required to carry part of the burden.” So too here, the FTCA’s retention of sovereign immunity for the Government’s discretionary acts does not imply a defense for the benefit of contractors who participate in those acts, even though they might pass on the financial burden to the United States. In either case, the most that can be said is that the position “asserted, though the product of a law Congress passed, is a matter on which Congress has not taken a position.”

Here, even that much is an overstatement, for the Government’s immunity for discretionary functions is not even “a product of” the FTCA. Before Congress enacted the FTCA (when sovereign immunity barred any tort suit against the Federal Government) we perceived no need for a rule of federal common law to reinforce the Government’s immunity by shielding also parties who might contractually pass costs on to it. Nor did we (or any other court of which I am aware) identify a special category of “discretionary” functions for which sovereign immunity was so crucial that a Government contractor who exercised discretion should share the Government’s immunity from state tort law.

Now, as before the FTCA’s enactment, the Federal Government is immune from “[a]ny claim . . . based upon the exercise or performance [of] a discretionary function,” including presumably any claim that petitioner might have brought against the Federal Government based upon respondent’s negligent design of the helicopter in which Lt.
Boyle died. There is no more reason for federal common law to shield contractors now that the Government is liable for some torts than there was when the Government was liable for none. The discretionary function exception does not support an immunity for the discretionary acts of Government contractors any more than the exception for “[a]ny claim [against the Government] arising out of assault,” § 2680(h), supports a personal immunity for Government employees who commit assaults. In short, while the Court purports to divine whether Congress would object to this suit, it inexplicably begins and ends its sortilege with an exception to a statute that is itself inapplicable and whose repeal would leave unchanged every relationship remotely relevant to the accident underlying this suit.

Far more indicative of Congress’ views on the subject is the wrongful-death cause of action that Congress itself has provided under the Death on the High Seas Act (DOHSA) (1920)—a cause of action that could have been asserted against United Technologies had Lt. Boyle’s helicopter crashed a mere three miles further off the coast of Virginia Beach. It is beyond me how a state-law tort suit against the designer of a military helicopter could be said to present any conflict, much less a “‘significant conflict,’” with “federal interests . . . in the context of Government procurement,” when federal law itself would provide a tort suit, but no (at least no explicit) Government-contractor defense, against the same designer for an accident involving the same equipment.

IV

At bottom, the Court’s analysis is premised on the proposition that any tort liability indirectly absorbed by the Government so burdens governmental functions as to compel us to act when Congress has not. That proposition is by no means uncontroversial. The tort system is premised on the assumption that the imposition of liability encourages actors to prevent any injury whose expected cost exceeds the cost of prevention. If the system is working as it should, Government contractors will design equipment to avoid certain injuries (like the deaths of soldiers or Government employees), which would be certain to burden the Government. The Court therefore has no basis for its assumption that tort liability will result in a net burden on the Government (let alone a clearly excessive net burden) rather than a net gain.

Perhaps tort liability is an inefficient means of ensuring the quality of design efforts, but “[w]hatever the merits of the policy” the Court wishes to implement, “its conversion into law is a proper subject for congressional action, not for any creative power of ours.” It is, after all, “Congress, not this Court or the other federal courts, [that] is the custodian of the national purse. By the same token [Congress] is the primary and most often the exclusive arbiter of federal fiscal affairs. And these comprehend, as we have said, securing the treasury or the Government against financial losses however inflicted . . . .” If Congress shared the Court’s assumptions and conclusion it could readily enact “A BILL [t]o place limitations on the civil liability of government contractors to ensure that such liability does not impede the ability of the United States to procure necessary goods and services,” H.R. 4765, 99th Cong., 2d Sess. (1986); see also S. 2441, 99th Cong., 2d Sess. (1986). It has not.
Were I a legislator, I would probably vote against any law absolving multibillion dollar private enterprises from answering for their tragic mistakes, at least if that law were justified by no more than the unsupported speculation that their liability might ultimately burden the United States Treasury. Some of my colleagues here would evidently vote otherwise (as they have here), but that should not matter here. We are judges not legislators, and the vote is not ours to cast . . . .

§ 7.3. The Dormant Commerce Clause and Flow Control


JUSTICE KENNEDY delivered the opinion of the Court.

... We consider a so-called flow control ordinance, which requires all solid waste to be processed at a designated transfer station before leaving the municipality. The avowed purpose of the ordinance is to retain the processing fees charged at the transfer station to amortize the cost of the facility. Because it attains this goal by depriving competitors, including out-of-state firms, of access to a local market, we hold that the flow control ordinance violates the Commerce Clause.

The town of Clarkstown, New York, ... agreed to close its landfill ... and build a new solid waste transfer station on the same site. The station would receive bulk solid waste and separate recyclable from nonrecyclable items. Recyclable waste would be baled for shipment to a recycling facility; nonrecyclable waste, to a suitable landfill or incinerator.

The cost of building the transfer station was estimated at $1.4 million. A local private contractor agreed to construct the facility and operate it for five years, after which the town would buy it for one dollar. During those five years, the town guaranteed a minimum waste flow of 120,000 tons per year, for which the contractor could charge the hauler a so-called tipping fee of $81 per ton. If the station received less than 120,000 tons in a year, the town promised to make up the tipping fee deficit. The object of this arrangement was to amortize the cost of the transfer station: the town would finance its new facility with the income generated by the tipping fees.

The problem, of course, was how to meet the yearly guarantee. This difficulty was compounded by the fact that the tipping fee of $81 per ton exceeded the disposal cost of unsorted solid waste on the private market. The solution the town adopted was the flow control ordinance here in question, Local Laws 1990, No. 9 of the Town of Clarkstown. The ordinance requires all nonhazardous solid waste within the town to be deposited at the ... transfer station ... .

... C & A Carbone, Inc. ... operates a recycling center in Clarkstown, where it receives bulk solid waste, sorts and bales it, and then ships it to other processing facilities—much as occurs at the town’s new transfer station. While the flow control ordinance permits recyclers like Carbone to continue receiving solid waste, it requires them to bring the
nonrecyclable residue from that waste to the [new] station. It thus forbids Carbone to ship the nonrecyclable waste itself, and it requires Carbone to pay a tipping fee on trash that Carbone has already sorted. . . .

At the outset, we confirm that the flow control ordinance does regulate interstate commerce, despite the town’s position to the contrary. The town says that its ordinance reaches only waste within its jurisdiction, and is, in practical effect, a quarantine: it prevents garbage from entering the stream of interstate commerce until it is made safe. This reasoning is premised, however, on an outdated and mistaken concept of what constitutes interstate commerce.

While the immediate effect of the ordinance is to direct local transport of solid waste to a designated site within the local jurisdiction, its economic effects are interstate in reach. The Carbone facility in Clarkstown receives and processes waste from places other than Clarkstown, including from out of State. By requiring Carbone to send the nonrecyclable portion of this waste to the . . . transfer station at an additional cost, the flow control ordinance drives up the cost for out-of-state interests to dispose of their solid waste. Furthermore, even as to waste originant in Clarkstown, the ordinance prevents everyone except the favored local operator from performing the initial processing step. The ordinance thus deprives out-of-state businesses of access to a local market. These economic effects are more than enough to bring the Clarkstown ordinance within the purview of the Commerce Clause. It is well settled that actions are within the domain of the Commerce Clause if they burden interstate commerce or impede its free flow.

The real question is whether the flow control ordinance is valid despite its undoubted effect on interstate commerce. For this inquiry, our case law yields two lines of analysis: first, whether the ordinance discriminates against interstate commerce; and second, whether the ordinance imposes a burden on interstate commerce that is “clearly excessive in relation to the putative local benefits” [(see Pike v. Bruce Church, Inc. (1970))]. As we find that the ordinance discriminates against interstate commerce, we need not resort to the Pike test.

The central rationale for the rule against discrimination is to prohibit state or municipal laws whose object is local economic protectionism, laws that would excite those jealousies and retaliatory measures the Constitution was designed to prevent. We have interpreted the Commerce Clause to invalidate local laws that impose commercial barriers or discriminate against an article of commerce by reason of its origin or destination out of State.

Clarkstown protests that its ordinance does not discriminate, because it does not differentiate solid waste on the basis of its geographic origin. All solid waste, regardless of origin, must be processed at the designated transfer station before it leaves the town. . . . [T]he ordinance erects no barrier to the import or export of any solid waste, but requires only that the waste be channeled through the designated facility.

Our initial discussion of the effects of the ordinance on interstate commerce goes far toward refuting the town’s contention that there is no discrimination in its regulatory
scheme. The town’s own arguments go the rest of the way. As the town itself points out, what makes garbage a profitable business is not its own worth but the fact that its possessor must pay to get rid of it. In other words, the article of commerce is not so much the solid waste itself, but rather the service of processing and disposing of it.

With respect to this stream of commerce, the flow control ordinance discriminates, for it allows only the favored operator to process waste that is within the limits of the town. The ordinance is no less discriminatory because in-state or in-town processors are also covered by the prohibition. In *Dean Milk Co. v. Madison* (1951), we struck down a city ordinance that required all milk sold in the city to be pasteurized within five miles of the city lines. We found it “immaterial that Wisconsin milk from outside the Madison area is subjected to the same proscription as that moving in interstate commerce.”

In this light, the flow control ordinance is just one more instance of local processing requirements that we long have held invalid. The essential vice in laws of this sort is that they bar the import of the processing service. Out-of-state meat inspectors, or shrimp hullers, or milk pasteurizers, are deprived of access to local demand for their services. Put another way, the offending local laws hoard a local resource—be it meat, shrimp, or milk—for the benefit of local businesses that treat it.

The flow control ordinance has the same design and effect. It hoards solid waste, and the demand to get rid of it, for the benefit of the preferred processing facility. The only conceivable distinction from the cases cited above is that the flow control ordinance favors a single local proprietor. But this difference just makes the protectionist effect of the ordinance more acute. In *Dean Milk*, the local processing requirement at least permitted pasteurizers within five miles of the city to compete. An out-of-state pasteurizer who wanted access to that market might have built a pasteurizing facility within the radius. The flow control ordinance at issue here squelches competition in the waste-processing service altogether, leaving no room for investment from outside.

Discrimination against interstate commerce in favor of local business or investment is per se invalid, save in a narrow class of cases in which the municipality can demonstrate, under rigorous scrutiny, that it has no other means to advance a legitimate local interest. A number of amici contend that the flow control ordinance fits into this narrow class. They suggest that as landfill space diminishes and environmental cleanup costs escalate, measures like flow control become necessary to ensure the safe handling and proper treatment of solid waste.

The teaching of our cases is that these arguments must be rejected absent the clearest showing that the unobstructed flow of interstate commerce itself is unable to solve the local problem. The Commerce Clause presumes a national market free from local legislation that discriminates in favor of local interests. Here Clarkstown has any number of nondiscriminatory alternatives for addressing the health and environmental problems alleged to justify the ordinance in question. The most obvious would be uniform safety regulations enacted without the object to discriminate. These regulations would ensure that competitors like Carbone do not underprice the market by cutting corners on environmental safety.
Nor may Clarkstown justify the flow control ordinance as a way to steer solid waste away from out-of-town disposal sites that it might deem harmful to the environment. To do so would extend the town’s police power beyond its jurisdictional bounds. States and localities may not attach restrictions to exports or imports in order to control commerce in other states.

The flow control ordinance does serve a central purpose that a nonprotectionist regulation would not: it ensures that the town-sponsored facility will be profitable, so that the local contractor can build it and Clarkstown can buy it back at nominal cost in five years. In other words, as the most candid of amici and even Clarkstown admit, the flow control ordinance is a financing measure. By itself, of course, revenue generation is not a local interest that can justify discrimination against interstate commerce. Otherwise, States could impose discriminatory taxes against solid waste originating outside the State.

Clarkstown maintains that special financing is necessary to ensure the long-term survival of the designated facility. If so, the town may subsidize the facility through general taxes or municipal bonds. But having elected to use the open market to earn revenues for its project, the town may not employ discriminatory regulation to give that project an advantage over rival businesses from out of State.

Though the Clarkstown ordinance may not in explicit terms seek to regulate interstate commerce, it does so nonetheless by its practical effect and design. In this respect the ordinance is not far different from the state law this Court found invalid in Buck v. Kuykendall (1925). That statute prohibited common carriers from using state highways over certain routes without a certificate of public convenience. Writing for the Court, Justice Brandeis said of the law: “Its primary purpose is not regulation with a view to safety or to conservation of the highways, but the prohibition of competition. It determines not the manner of use, but the persons by whom the highways may be used. It prohibits such use to some persons while permitting it to others for the same purpose and in the same manner.”

State and local governments may not use their regulatory power to favor local enterprise by prohibiting patronage of out-of-state competitors or their facilities.

**Justice O’Connor, concurring in the judgment.**

The town of Clarkstown’s flow control ordinance requires all “acceptable waste” generated or collected in the town to be disposed of only at the town’s solid waste facility. The Court holds today that this ordinance violates the Commerce Clause because it discriminates against interstate commerce. I agree with the majority’s ultimate conclusion that the ordinance violates the dormant Commerce Clause. In my view, however, the town’s ordinance is unconstitutional not because of facial or effective discrimination against interstate commerce, but rather because it imposes an excessive burden on interstate commerce [under the Pike test].
The scope of the dormant Commerce Clause is a judicial creation. On its face, the Clause provides only that “[t]he Congress shall have Power . . . To regulate Commerce . . . among the several States. . . .” This Court long ago concluded, however, . . . that the dormant Commerce Clause forbids States and their subdivisions from regulating interstate commerce.

We have generally distinguished between two types of impermissible regulations. A facially nondiscriminatory regulation supported by a legitimate state interest which incidentally burdens interstate commerce is constitutional unless the burden on interstate trade is clearly excessive in relation to the local benefits. Where, however, a regulation “affirmatively” or “clearly” discriminates against interstate commerce on its face or in practical effect, it violates the Constitution unless the discrimination is demonstrably justified by a valid factor unrelated to protectionism. Of course, there is no clear line separating these categories. “In either situation the critical consideration is the overall effect of the statute on both local and interstate activity.”

Local Law 9 prohibits anyone except the town-authorized transfer station operator from processing discarded waste and shipping it out of town. In effect, the town has given a waste processing monopoly to the transfer station. The majority concludes that this processing monopoly facially discriminates against interstate commerce. . . .

Local Law 9, however, lacks an important feature common to the regulations at issue in [past] cases—namely, discrimination on the basis of geographic origin. In each of the cited cases, the challenged enactment gave a competitive advantage to local business as a group vis-a-vis their out-of-state or nonlocal competitors as a group. In effect, the regulating jurisdiction—be it a State, a county, or a city—drew a line around itself and treated those inside the line more favorably than those outside the line. Thus, in *Pike*, the Court held that an Arizona law requiring that Arizona cantaloupes be packaged in Arizona before being shipped out of state facially discriminated against interstate commerce: the benefits of the discriminatory scheme benefited the Arizona packaging industry, at the expense of its competition in California. Similarly, in *Dean Milk*, on which the majority heavily relies, the city of Madison drew a line around its perimeter and required that all milk sold in the City be pasteurized only by dairies located inside the line. This type of geographic distinction, which confers an economic advantage on local interests in general, is common to all the local processing cases cited by the majority. And the Court has, I believe, correctly concluded that these arrangements are protectionist either in purpose or practical effect, and thus amount to virtually per se discrimination.

. . . Unlike the regulations we have previously struck down, Local Law 9 does not give more favorable treatment to local interests as a group as compared to out-of-state or out-of-town economic interests. Rather, the garbage sorting monopoly is achieved at the expense of all competitors, be they local or nonlocal. That the ordinance does not discriminate on the basis of geographic origin is vividly illustrated by the identity of the plaintiff in this very action: petitioner is a local recycler, physically located in
Clarkstown, that desires to process waste itself, and thus bypass the town’s designated transfer facility. Because in-town processors—like petitioner—and out-of-town processors are treated equally, I cannot agree that Local Law 9 “discriminates” against interstate commerce. Rather, Local Law 9 “discriminates” evenhandedly against all potential participants in the waste processing business, while benefiting only the chosen operator of the transfer facility.

I believe this distinction has more doctrinal significance than the majority acknowledges. In considering state health and safety regulations such as Local Law 9, we have consistently recognized that the fact that interests within the regulating jurisdiction are equally affected by the challenged enactment counsels against a finding of discrimination. And for good reason. The existence of substantial in-state interests harmed by a regulation is “a powerful safeguard” against legislative discrimination. The Court generally defers to health and safety regulations because “their burden usually falls on local economic interests, as well as other States’ economic interests, thus insuring that a State’s own political processes will serve as a check against unduly burdensome regulations.” Thus, while there is no bright line separating those enactments which are virtually per se invalid and those which are not, the fact that in-town competitors of the transfer facility are equally burdened by Local Law 9 leads me to conclude that Local Law 9 does not discriminate against interstate commerce.

II

That the ordinance does not discriminate against interstate commerce does not, however, end the Commerce Clause inquiry. Even a nondiscriminatory regulation may nonetheless impose an excessive burden on interstate trade when considered in relation to the local benefits conferred. Indeed, we have long recognized that “a burden imposed by a State upon interstate commerce is not to be sustained simply because the statute imposing it applies alike to . . . the people of the State enacting such statute.” Moreover, “the extent of the burden that will be tolerated will, of course, depend on the nature of the local interest involved, and on whether it could be promoted as well with a lesser impact on interstate activities.” Judged against these standards, Local Law 9 fails.

The local interest in proper disposal of waste is obviously significant. But this interest could be achieved by simply requiring that all waste disposed of in the town be properly processed somewhere. For example, the town could ensure proper processing by setting specific standards with which all town processors must comply.

In fact, however, the town’s purpose is narrower than merely ensuring proper disposal. Local Law 9 is intended to ensure the financial viability of the transfer facility. I agree with the majority that this purpose can be achieved by other means that would have a less dramatic impact on the flow of goods. For example, the town could finance the project by imposing taxes, by issuing municipal bonds, or even by lowering its price for processing to a level competitive with other waste processing facilities. But by requiring that all waste be processed at the town’s facility, the ordinance “squelches competition in the waste-processing service altogether, leaving no room for investment from outside.”
In addition, “[t]he practical effect of [Local Law 9] must be evaluated not only by considering the consequences of the statute itself, but also by considering how the challenged statute may interact with the legitimate regulatory regimes of the other States and what effect would arise if not one, but many or every, [jurisdiction] adopted similar legislation.” This is not a hypothetical inquiry. Over 20 states have enacted statutes authorizing local governments to adopt flow control laws. If the localities in these States impose the type of restriction on the movement of waste that Clarkstown has adopted, the free movement of solid waste in the stream of commerce will be severely impaired. Indeed, pervasive flow control would result in the type of balkanization the Clause is primarily intended to prevent.

Given that many jurisdictions are contemplating or enacting flow control, the potential for conflicts is high. For example, in the State of New Jersey, just south of Clarkstown, local waste may be removed from the State for the sorting of recyclables “as long as the residual solid waste is returned to New Jersey.” Under Local Law 9, however, if petitioners bring waste from New Jersey for recycling at their Clarkstown operation, the residual waste may not be returned to New Jersey, but must be transported to Clarkstown’s transfer facility. As a consequence, operations like petitioners’ cannot comply with the requirements of both jurisdictions. Nondiscriminatory state or local laws which actually conflict with the enactments of other States are constitutionally infirm if they burden interstate commerce. The increasing number of flow control regimes virtually ensures some inconsistency between jurisdictions, with the effect of eliminating the movement of waste between jurisdictions. I therefore conclude that the burden Local Law 9 imposes on interstate commerce is excessive in relation to Clarkstown’s interest in ensuring a fixed supply of waste to supply its project. . . .

Justice SOUTER, with whom THE CHIEF JUSTICE and Justice BLACKMUN join, dissenting.

The majority may invoke “well-settled principles of our Commerce Clause jurisprudence,” but it does so to strike down an ordinance unlike anything this Court has ever invalidated. Previous cases have held that the “negative” or “dormant” aspect of the Commerce Clause renders state or local legislation unconstitutional when it discriminates against out-of-state or out-of-town businesses such as those that pasteurize milk, hull shrimp, or mill lumber, and the majority relies on these cases because of what they have in common with this one: out-of-state processors are excluded from the local market (here, from the market for trash processing services). What the majority ignores, however, are the differences between our local processing cases and this one: the exclusion worked by Clarkstown’s Local Law 9 bestows no benefit on a class of local private actors, but instead directly aids the government in satisfying a traditional governmental responsibility. The law does not differentiate between all local and all out-of-town providers of a service, but instead between the one entity responsible for ensuring that the job gets done and all other enterprises, regardless of their location. The ordinance thus falls outside that class of tariff or protectionist measures that the Commerce Clause has traditionally been thought to bar States from enacting against each other, and when the majority subsumes the ordinance within the class of laws this Court has struck down as facially discriminatory (and so avails itself of our “virtually per se
rule” against such statutes), the majority is in fact greatly extending the Clause’s dormant reach.

There are, however, good and sufficient reasons against expanding the Commerce Clause’s inherent capacity to trump exercises of state authority such as the ordinance at issue here. There is no indication in the record that any out-of-state trash processor has been harmed, or that the interstate movement or disposition of trash will be affected one whit. To the degree Local Law 9 affects the market for trash processing services, it does so only by subjecting Clarkstown residents and businesses to burdens far different from the burdens of local favoritism that dormant Commerce Clause jurisprudence seeks to root out. The town has found a way to finance a public improvement, not by transferring its cost to out-of-state economic interests, but by spreading it among the local generators of trash, an equitable result with tendencies that should not disturb the Commerce Clause and should not be disturbed by us.

II

Carbone’s complaint is one that any Clarkstown trash generator could have made: the town has created a monopoly on trash processing services, and residents are no longer free to provide these services for themselves or to contract for them with others at a mutually agreeable price. *

We are not called upon to judge the ultimate wisdom of creating this local monopoly, but we are asked to say whether Clarkstown’s monopoly violates the Commerce Clause, as long read by this Court to limit the power of state and local governments to discriminate against interstate commerce . . . .

This limitation on the state and local power has been seen implicit in the Commerce Clause because, as the majority recognizes, the Framers sought to dampen regional jealousies in general and, in particular, to eliminate retaliatory tariffs, which had poisoned commercial relations under the Articles of Confederation. Laws that hoard for local businesses the right to serve local markets or develop local resources work to isolate States from each other and to incite retaliation, since no State would stand by while another advanced the economic interests of its own business classes at the expense of its neighbors.

A

The majority argues that resolution of the issue before us is controlled by a line of cases in which we have struck down state or local laws that discriminate against out-of-state or out-of-town providers of processing services. With perhaps one exception, the laws invalidated in those cases were patently discriminatory, differentiating by their very terms between in-state and out-of-state (or local and nonlocal) processors. One ordinance, for example, forbid selling pasteurized milk “‘unless the same shall have been pasteurized and bottled . . . within a radius of five miles from the central portion of the City of Madison. . . .’” The other laws expressly discriminated against commerce crossing state

* This sentence was moved from section I.
lines, placing these local processing cases squarely within the larger class of cases in which this Court has invalidated facially discriminatory legislation.

As the majority recognizes, Local Law 9 shares two features with these local processing cases. It regulates a processing service available in interstate commerce, i.e., the sorting and baling of solid waste for disposal. And it does so in a fashion that excludes out-of-town trash processors by its very terms. These parallels between Local Law 9 and the statutes previously invalidated confer initial plausibility on the majority’s classification of this case with those earlier ones on processing, and they even bring this one within the most general language of some of the earlier cases, abhorring the tendency of such statutes “to impose an artificial rigidity on the economic pattern of the industry.”

B

. . . [H]owever, . . . the terms of Clarkstown’s ordinance favor a single processor, not the class of all such businesses located in Clarkstown. Second, the one proprietor so favored is essentially an agent of the municipal government, which (unlike Carbone or other private trash processors) must ensure the removal of waste according to acceptable standards of public health. Any discrimination worked by Local Law 9 thus fails to produce the sort of entrepreneurial favoritism we have previously defined and condemned as protectionist.

I

The outstanding feature of the statutes reviewed in the local processing cases is their distinction between two classes of private economic actors according to location, favoring shrimp hullers within Louisiana, milk pasteurizers within five miles of the center of Madison, and so on. Since nothing in these local processing laws prevented a proliferation of local businesses within the State or town, the out-of-town processors were not excluded as part and parcel of a general exclusion of private firms from the market, but as a result of discrimination among such firms according to geography alone. It was because of that discrimination in favor of local businesses, preferred at the expense of their out-of-town or out-of-state competitors, that the Court struck down those local processing laws as classic examples of the economic protectionism the dormant Commerce Clause jurisprudence aims to prevent. In the words of one commentator summarizing our case law, it is laws “adopted for the purpose of improving the competitive position of local economic actors, just because they are local, vis-a-vis their foreign competitors” that offend the Commerce Clause. The Commerce Clause does not otherwise protect access to local markets.

The majority recognizes, but discounts, this difference between laws favoring all local actors and this law favoring a single municipal one. According to the majority, “this difference just makes the protectionist effect of the ordinance more acute,” because outside investors cannot even build competing facilities within Clarkstown. But of course Clarkstown investors face the same prohibition, which is to say that Local Law 9’s exclusion of outside capital is part of a broader exclusion of private capital, not a discrimination against out-of-state investors as such. Thus, while these differences may

502
Draft — do not distribute!
underscore the ordinance’s anticompetitive effect, they substantially mitigate any protectionist effect, for subjecting out-of-town investors and facilities to the same constraints as local ones is not economic protectionism.

Nor is the monopolist created by Local Law 9 just another private company successfully enlisting local government to protect the jobs and profits of local citizens. While our previous local processing cases have barred discrimination in markets served by private companies, Clarkstown’s transfer station is essentially a municipal facility, built and operated under a contract with the municipality and soon to revert entirely to municipal ownership. This, of course, is no mere coincidence, since the facility performs a municipal function that tradition as well as state and federal law recognize as the domain of local government. Throughout the history of this country, municipalities have taken responsibility for disposing of local garbage to prevent noisome smells, obstruction of the streets, and threats to public health, and today 78 percent of landfills receiving municipal solid waste are owned by local governments. The National Government provides “technical and financial assistance to States or regional authorities for comprehensive planning” with regard to the disposal of solid waste, and the State of New York authorizes local governments to prepare such management plans for the proper disposal of all solid waste generated within their jurisdictions. These general provisions underlie Clarkstown’s more specific obligation (under its consent decree with the New York State Department of Environmental Conservation) to establish a transfer station in place of the old town dump, and it is to finance this transfer station that Local Law 9 was passed.

The majority ignores this distinction between public and private enterprise, equating Local Law 9’s “hoard[ing]” of solid waste for the municipal transfer station with the design and effect of ordinances that restrict access to local markets for the benefit of local private firms. But private businesses, whether local or out of State, first serve the private interests of their owners, and there is therefore only rarely a reason other than economic protectionism for favoring local businesses over their out-of-town competitors. The local government itself occupies a very different market position, however, being the one entity that enters the market to serve the public interest of local citizens quite apart from private interest in private gain. Reasons other than economic protectionism are accordingly more likely to explain the design and effect of an ordinance that favors a public facility. The facility as constructed might, for example, be one that private economic actors, left to their own devices, would not have built, but which the locality needs in order to abate (or guarantee against creating) a public nuisance. There is some evidence in this case that this is so, as the New York State Department of Environmental Conservation would have had no reason to insist that Clarkstown build its own transfer station if the private market had furnished adequate processing capacity to meet Clarkstown’s needs. An ordinance that favors a municipal facility, in any event, is one that favors the public sector, and if “we continue to recognize that the States occupy a special and specific position in our constitutional system and that the scope of Congress’ authority under the Commerce Clause must reflect that position,” then surely this Court’s dormant Commerce Clause jurisprudence must itself see that favoring state-sponsored
facilities differs from discriminating among private economic actors, and is much less likely to be protectionist.

Having established that Local Law 9 does not serve the competitive class identified in previous local processing cases and that Clarkstown differs correspondingly from other local processors, we must ask whether these differences justify a standard of dormant Commerce Clause review that differs from the virtually fatal scrutiny imposed in those earlier cases. I believe they do.

The justification for subjecting the local processing laws and the broader class of clearly discriminatory commercial regulation to near-fatal scrutiny is the virtual certainty that such laws, at least in their discriminatory aspect, serve no legitimate, nonprotectionist purpose. Whether we find the “the evil of protectionism” in the clear import of specific statutory provisions or in the legislature’s ultimate purpose, the discriminatory scheme is almost always designed either to favor local industry, as such, or to achieve some other goal while exporting a disproportionate share of the burden of attaining it, which is merely a subtler form of local favoritism.

On the other hand, in a market served by a municipal facility, a law that favors that single facility over all others is a law that favors the public sector over all private sector processors, whether local or out of State. Because the favor does not go to local private competitors of out-of-state firms, out-of-state governments will at the least lack a motive to favor their own firms in order to equalize the positions of private competitors. While a preference in favor of the government may incidentally function as local favoritism as well, a more particularized enquiry is necessary before a court can say whether such a law does in fact smack too strongly of economic protectionism. If Local Law 9 is to be struck down, in other words, it must be under that test most readily identified with *Pike*.

We have said that when legislation that does not facially discriminate “comes into conflict with the Commerce Clause’s overriding requirement of a national “common market,” we are confronted with the task of effecting an accommodation of the competing national and local interests.” Although this analysis of competing interests has sometimes been called a “balancing test,” it is not so much an open-ended weighing of an ordinance’s pros and cons as an assessment of whether an ordinance discriminates in practice, or otherwise unjustifiably operates to isolate a State’s economy from the national common market. If a statute or local ordinance serves a legitimate local interest and does not patently discriminate, “it will be upheld unless the burden imposed on [interstate] commerce is clearly excessive in relation to the putative local benefits.” The analysis is similar to, but softer around the edges than, the test we employ in cases of overt discrimination. “[T]he question becomes one of degree,” and its answer depends on the nature of the burden on interstate commerce, the nature of the local interest, and the availability of alternative methods for advancing the local interest without hindering the national one.
The primary burden Carbone attributes to flow control ordinances such as Local Law 9 is that they “prevent trash from being sent to the most cost-effective disposal facilities, and insulate the designated facility from all price competition.” In this case, customers must pay $11 per ton more for dumping trash at the Clarkstown transfer station than they would pay at Carbone’s facility . . . .

Fortunately, the dollar cost of the burden need not be pinpointed, its nature being more significant than its economic extent. When we look to its nature, it should be clear that the monopolistic character of Local Law 9’s effects is not itself suspicious for purposes of the Commerce Clause. Although the right to compete is a hallmark of the American economy, and local monopolies are subject to challenge under the century-old Sherman Act, the bar to monopolies (or, rather, the authority to dismember and penalize them) arises from a statutory, not a constitutional, mandate. No more than the Fourteenth Amendment, the Commerce Clause “does not enact Mr. Herbert Spencer’s Social Statics . . . [or] embody a particular economic theory, whether of paternalism . . . or of laissez faire.” *Lochner v. New York* (1905) (Holmes, J., dissenting). The dormant Commerce Clause does not “protec[t] the particular structure or methods of operation in a[ny] . . . market.” The only right to compete that it protects is the right to compete on terms independent of one’s location.

While the monopolistic nature of the burden may be disregarded, any geographically discriminatory elements must be assessed with care. We have already observed that there is no geographically based selection among private firms, and it is clear from the face of the ordinance that nothing hinges on the source of trash that enters Clarkstown or upon the destination of the processed waste that leaves the transfer station. There is, to be sure, an incidental local economic benefit, for the need to process Clarkstown’s trash in Clarkstown will create local jobs. But this local boon is mitigated by another feature of the ordinance, in that it finances whatever benefits it confers on the town from the pockets of the very citizens who passed it into law. On the reasonable assumption that no one can avoid producing some trash, every resident of Clarkstown must bear a portion of the burden Local Law 9 imposes to support the municipal monopoly, an uncharacteristic feature of statutes claimed to violate the Commerce Clause.

By way of contrast, most of the local processing statutes we have previously invalidated imposed requirements that made local goods more expensive as they headed into the national market, so that out-of-state economies bore the bulk of any burden. Requiring that Alaskan timber be milled in that State prior to export would add the value of the milling service to the Alaskan economy at the expense of some other State, but would not burden the Alaskans who adopted such a law. Similarly, South Carolinians would retain the financial benefit of a local processing requirement for shrimp without paying anything more themselves. And in *Philadelphia v. New Jersey*, the State attempted to export the burden of conserving its scarce landfill space by barring the importation of out-of-state waste. Courts step in through the dormant Commerce Clause to prevent such exports because legislative action imposing a burden “‘principally upon those without the state . . . is not likely to be subjected to those political restraints which are normally exerted on legislation where it affects adversely some interests within the state.’” Here, in contrast, every voter in Clarkstown pays to fund the benefits of flow control, however
high the tipping fee is set. Since, indeed, the mandate to use the town facility will only make a difference when the tipping fee raises the cost of using the facility above what the market would otherwise set, the Clarkstown voters are funding their benefit by assessing themselves and paying an economic penalty. Any whiff of economic protectionism is far from obvious. . . .

Moreover, flow control offers an additional benefit that could not be gained by financing through a subsidy derived from general tax revenues, in spreading the cost of the facility among all Clarkstown residents who generate trash. The ordinance does, of course, protect taxpayers, including those who already support the transfer station by patronizing it, from ending up with the tab for making provision for large-volume trash producers like Carbone, who would rely on the municipal facility when that was advantageous but opt out whenever the transfer station’s price rose above the market price. In proportioning each resident’s burden to the amount of trash generated, the ordinance has the added virtue of providing a direct and measurable deterrent to the generation of unnecessary waste in the first place. And in any event it is far from clear that the alternative to flow control (i.e., subsidies from general tax revenues or municipal bonds) would be less disruptive of interstate commerce than flow control, since a subsidized competitor can effectively squelch competition by underbidding it. . . .

§ 7.3.2. United Haulers Ass’n, Inc. v. Oneida-Herkimer Solid Waste Mgmt. Auth., 127 S. Ct. 1786 (2007)

CHIEF JUSTICE ROBERTS delivered the opinion of the Court.*

. . . In this case, we face flow control ordinances quite similar to the one invalidated in Carbone. The only salient difference is that the laws at issue here require haulers to bring waste to facilities owned and operated by a state-created public benefit corporation. We find this difference constitutionally significant. Disposing of trash has been a traditional government activity for years, and laws that favor the government in such areas—but treat every private business, whether in-state or out-of-state, exactly the same—do not discriminate against interstate commerce for purposes of the Commerce Clause. Applying the Commerce Clause test reserved for regulations that do not discriminate against interstate commerce, we uphold these ordinances because any incidental burden they may have on interstate commerce does not outweigh the benefits they confer on the citizens of Oneida and Herkimer Counties. . . .

* Except as otherwise indicated.
The flow control ordinances in this case benefit a clearly public facility, while treating all private companies exactly the same. . . . [S]uch flow control ordinances do not discriminate against interstate commerce for purposes of the dormant Commerce Clause.

Compelling reasons justify treating these laws differently from laws favoring particular private businesses over their competitors. “Conceptually, of course, any notion of discrimination assumes a comparison of substantially similar entities.” But States and municipalities are not private businesses—far from it. Unlike private enterprise, government is vested with the responsibility of protecting the health, safety, and welfare of its citizens. These important responsibilities set state and local government apart from a typical private business.

Given these differences, it does not make sense to regard laws favoring local government and laws favoring private industry with equal skepticism. As our local processing cases demonstrate, when a law favors in-state business over out-of-state competition, rigorous scrutiny is appropriate because the law is often the product of “simple economic protectionism.” Laws favoring local government, by contrast, may be directed toward any number of legitimate goals unrelated to protectionism. Here the flow control ordinances enable the Counties to pursue particular policies with respect to the handling and treatment of waste generated in the Counties, while allocating the costs of those policies on citizens and businesses according to the volume of waste they generate.

The contrary approach of treating public and private entities the same under the dormant Commerce Clause would lead to unprecedented and unbounded interference by the courts with state and local government. The dormant Commerce Clause is not a roving license for federal courts to decide what activities are appropriate for state and local government to undertake, and what activities must be the province of private market competition. In this case, the citizens of Oneida and Herkimer Counties have chosen the government to provide waste management services, with a limited role for the private sector in arranging for transport of waste from the curb to the public facilities. The citizens could have left the entire matter for the private sector, in which case any regulation they undertook could not discriminate against interstate commerce. But it was also open to them to vest responsibility for the matter with their government, and to adopt flow control ordinances to support the government effort. It is not the office of the Commerce Clause to control the decision of the voters on whether government or the private sector should provide waste management services. “The Commerce Clause significantly limits the ability of States and localities to regulate or otherwise burden the flow of interstate commerce, but it does not elevate free trade above all other values.”

We should be particularly hesitant to interfere with the Counties’ efforts under the guise of the Commerce Clause because “[w]aste disposal is both typically and traditionally a local government function.” Congress itself has recognized local government’s vital role in waste management, making clear that “collection and disposal of solid wastes should
continue to be primarily the function of State, regional, and local agencies.” The policy of the State of New York favors “displace[ing] competition with regulation or monopoly control” in this area. We may or may not agree with that approach, but nothing in the Commerce Clause vests the responsibility for that policy judgment with the Federal Judiciary.

Finally, it bears mentioning that the most palpable harm imposed by the ordinances—more expensive trash removal—is likely to fall upon the very people who voted for the laws. Our dormant Commerce Clause cases often find discrimination when a State shifts the costs of regulation to other States, because when “the burden of state regulation falls on interests outside the state, it is unlikely to be alleviated by the operation of those political restraints normally exerted when interests within the state are affected.” Here, the citizens and businesses of the Counties bear the costs of the ordinances. There is no reason to step in and hand local businesses a victory they could not obtain through the political process.

We hold that the Counties’ flow control ordinances, which treat in-state private business interests exactly the same as out-of-state ones, do not “discriminate against interstate commerce” for purposes of the dormant Commerce Clause. [Later in the opinion, Chief Justice Roberts—no longer writing for a majority of the Court—finds that the ordinances are valid under the Pike balancing test.] . . .

Justice ALITO, with whom Justice STEVENS and Justice KENNEDY join, dissenting.

I

. . . This case cannot be meaningfully distinguished from Carbone. As the Court itself acknowledges, “[t]he only salient difference” between the cases is that the ordinance invalidated in Carbone discriminated in favor of a privately owned facility, whereas the laws at issue here discriminate in favor of “facilities owned and operated by a state-created public benefit corporation.” The Court relies on the distinction between public and private ownership to uphold the flow-control laws, even though a straightforward application of Carbone would lead to the opposite result. The public-private distinction drawn by the Court is both illusory and without precedent.

II

The fact that the flow control laws at issue discriminate in favor of a government-owned enterprise does not meaningfully distinguish this case from Carbone. The preferred facility in Carbone was, to be sure, nominally owned by a private contractor who had built the facility on the town’s behalf, but it would be misleading to describe the facility as private. In exchange for the contractor’s promise to build the facility for the town free of charge and then to sell it to the town five years later for $1, the town guaranteed that, during the first five years of the facility’s existence, the contractor would receive “a minimum waste flow of 120,000 tons per year” and that the contractor could charge an above-market tipping fee. If the facility “received less than 120,000 tons in a year, the
town [would] make up the tipping fee deficit.” To prevent residents, businesses, and trash haulers from taking their waste elsewhere in pursuit of lower tipping fees (leaving the town responsible for covering any shortfall in the contractor’s guaranteed revenue stream), the town enacted an ordinance “requir[ing] all nonhazardous solid waste within the town to be deposited at” the preferred facility.

This Court observed that “[t]he object of this arrangement was to amortize the cost of the transfer station: The town would finance its new facility with the income generated by the tipping fees.” “In other words,” the Court explained, “the flow control ordinance [wa]s a financing measure,” for what everyone—including the Court—regarded as the town’s new transfer station.

The only real difference between the facility at issue in Carbone and its counterpart in this case is that title to the former had not yet formally passed to the municipality. The Court exalts form over substance in adopting a test that turns on this technical distinction, particularly since, barring any obstacle presented by state law, the transaction in Carbone could have been restructured to provide for the passage of title at the beginning, rather than the end, of the 5-year period.

For this very reason, it is not surprising that in Carbone the Court did not dispute the dissent’s observation that the preferred facility was for all practical purposes owned by the municipality. To the contrary, the Court repeatedly referred to the transfer station in terms suggesting that the transfer station did in fact belong to the town.

Today the Court dismisses those statements as “at best inconclusive.” The Court, however, fails to offer any explanation as to what other meaning could possibly attach to Carbone’s repeated references to Clarkstown’s transfer station as a municipal facility. It also ignores the fact that the ordinance itself, which was included in its entirety in an appendix to the Court’s opinion, repeatedly referred to the station as “the Town of Clarkstown solid waste facility.” The Court likewise fails to acknowledge that the parties in Carbone openly acknowledged the municipal character of the transfer station. . . .

III

In any event, we have never treated discriminatory legislation with greater deference simply because the entity favored by that legislation was a government-owned enterprise. . . . The Court has long subjected discriminatory legislation to strict scrutiny, and has never, until today, recognized an exception for discrimination in favor of a state-owned entity.

A

This Court long ago recognized that the Commerce Clause can be violated by a law that discriminates in favor of a state-owned monopoly. In the 1890’s, South Carolina enacted laws giving a state agency the exclusive right to operate facilities selling alcoholic beverages within that State, and these laws were challenged under the Commerce Clause in Scott v. Donald (1897) and Vance v. W. A. Vandercook Co. (1898). The Court held that the Commerce Clause barred the State from prohibiting its residents from purchasing
alcohol from out-of-state vendors, but that the State could surmount this problem by allowing residents to receive out-of-state shipments for their personal use. The Court’s holding was based on the same fundamental dormant Commerce Clause principle applied in *Carbone*. As the Court put it in *Vance*, a State “‘cannot discriminate against the bringing of [lawful] articles in and importing them from other States’’” because such discrimination is “‘a hindrance to interstate commerce and an unjust preference of the products of the enacting State as against similar products of other States.’”

Thus, were it not for the Twenty-first Amendment, laws creating state-owned liquor monopolies—which many States maintain today—would be deemed discriminatory under the dormant Commerce Clause. There is, of course, no comparable provision in the Constitution authorizing States to discriminate against out-of-state providers of waste processing and disposal services, either by means of a government-owned monopoly or otherwise.

IV

Despite precedent condemning discrimination in favor of government-owned enterprises, the Court attempts to develop a logical justification for the rule it creates today. That justification rests on three principal assertions. First, the Court insists that it simply “does not make sense to regard laws favoring local government and laws favoring private industry with equal skepticism,” because the latter are “often the product of ‘simple economic protectionism,’” while the former “may be directed toward any number of legitimate goals unrelated to protectionism.” Second, the Court reasons that deference to legislation discriminating in favor of a municipal landfill is especially appropriate considering that “[w]aste disposal is both typically and traditionally a local government function.” Third, the Court suggests that respondents’ flow-control laws are not discriminatory because they “treat in-state private business interests exactly the same as out-of-state ones.” I find each of these arguments unpersuasive.

A

I see no basis for the Court’s assumption that discrimination in favor of an in-state facility owned by the government is likely to serve “legitimate goals unrelated to protectionism.” Discrimination in favor of an in-state government facility serves “‘local economic interests,’” inuring to the benefit of local residents who are employed at the facility, local businesses that supply the facility with goods and services, and local workers employed by such businesses. It is therefore surprising to read in the opinion of the Court that state discrimination in favor of a state-owned business is not likely to be motivated by economic protectionism.

Experience in other countries, where state ownership is more common than it is in this country, teaches that governments often discriminate in favor of state-owned businesses (by shielding them from international competition) precisely for the purpose of protecting those who derive economic benefits from those businesses, including their employees. Such discrimination amounts to economic protectionism in any realistic sense of the term.
By the same token, discrimination in favor of an in-state, privately owned facility may serve legitimate ends, such as the promotion of public health and safety. For example, a State might enact legislation discriminating in favor of produce or livestock grown within the State, reasoning that the State’s inspectors can more easily monitor the use of pesticides, fertilizers, and feed on farms within the State’s borders. Such legislation would almost certainly be unconstitutional, notwithstanding its potential to promote public health and safety.

The fallacy in the Court’s approach can be illustrated by comparing a law that discriminates in favor of an in-state facility, owned by a corporation whose shares are publicly held, and a law discriminating in favor of an otherwise identical facility that is owned by the State or municipality. Those who are favored and disfavored by these two laws are essentially the same with one major exception: The law favoring the corporate facility presumably benefits the corporation’s shareholders, most of whom are probably not local residents, whereas the law favoring the government-owned facility presumably benefits the people of the enacting State or municipality. I cannot understand why only the former law, and not the latter, should be regarded as a tool of economic protectionism. Nor do I think it is realistic or consistent with our precedents to condemn some discriminatory laws as protectionist while upholding other, equally discriminatory laws as lawful measures designed to serve legitimate local interests unrelated to protectionism.

For these reasons, I cannot accept the proposition that laws discriminating in favor of state-owned enterprises are so unlikely to be the product of economic protectionism that they should be exempt from the usual dormant Commerce Clause standards.

Proper analysis under the dormant Commerce Clause involves more than an inquiry into whether the challenged Act is in some sense “directed toward . . . legitimate goals unrelated to protectionism”; equally important are the means by which those goals are realized. If the chosen means take the form of a statute that discriminates against interstate commerce—“‘either on its face or in practical effect’”—then “the burden falls on [the enacting government] to demonstrate both that the statute ‘serves a legitimate local purpose,’ and that this purpose could not be served as well by available nondiscriminatory means.”

Thus, if the legislative means are themselves discriminatory, then regardless of how legitimate and nonprotectionist the underlying legislative goals may be, the legislation is subject to strict scrutiny. Similarly, the fact that a discriminatory law “may [in some sense] be directed toward any number of legitimate goals unrelated to protectionism” does not make the law nondiscriminatory. The existence of such goals is relevant, not to whether the law is discriminatory, but to whether the law can be allowed to stand even though it discriminates against interstate commerce. And even then, the existence of legitimate goals is not enough; discriminatory legislation can be upheld only where such goals cannot adequately be achieved through nondiscriminatory means.

*Dean Milk* is instructive on this point. That case involved a dormant Commerce Clause challenge to an ordinance requiring all milk sold in Madison, Wisconsin, to be processed
within five miles of the city’s central square. The ordinance “professe[d] to be a health measure,” and may have conferred some benefit on the city and its residents to the extent that it succeeded in guaranteeing the purity and quality of the milk sold in the city. The Court nevertheless invalidated the ordinance, concluding that any public health benefits it may have conferred could be achieved through “reasonable nondiscriminatory alternatives,” including a system that would allow a nonlocal dairy to qualify to sell milk in the city upon proving that it was in compliance with applicable health and safety requirements.

The Court did not inquire whether the real purpose of the ordinance was to benefit public health and safety or to protect local economic interests; nor did the Court make any effort to determine whether or to what extent the ordinance may have succeeded in promoting health and safety. In fact, the Court apparently assumed that the ordinance could fairly be characterized as “a health measure.” The Court nevertheless concluded that the ordinance could not stand because it “erect[ed] an economic barrier protecting a major local industry against competition from without the State,” “plac[ed] a discriminatory burden on interstate commerce,” and was “not essential for the protection of local health interests.”

The overarching concern expressed by the Court was that the ordinance, if left intact, “would invite a multiplication of preferential trade areas destructive of the very purpose of the Commerce Clause.” “Under the circumstances here presented,” the Court concluded, “the regulation must yield to the principle that ‘one state in its dealings with another may not place itself in a position of economic isolation.’”

The same reasoning dooms the laws challenged here. Like the ordinance in Dean Milk, these laws discriminate against interstate commerce (generally favoring local interests over nonlocal interests), but are defended on the ground that they serve legitimate goals unrelated to protectionism (e.g., health, safety, and protection of the environment). And while I do not question that the laws at issue in this case serve legitimate goals, the laws offend the dormant Commerce Clause because those goals could be attained effectively through nondiscriminatory means. Indeed, no less than in Carbone, those goals could be achieved through “uniform [health and] safety regulations enacted without the object to discriminate” that “would ensure that competitors [to the municipal program] do not underprice the market by cutting corners on environmental safety.” Respondents would also be free, of course, to “subsidize the[ir] [program] through general taxes or municipal bonds.” “But having elected to use the open market to earn revenues for” their waste management program, respondents “may not employ discriminatory regulation to give that [program] an advantage over rival businesses from out of State.”

B

The Court next suggests that deference to legislation discriminating in favor of a municipal landfill is especially appropriate considering that “[w]aste disposal is both typically and traditionally a local government function.” I disagree on two grounds.
First, this Court has previously recognized that any standard “that turns on a judicial appraisal of whether a particular governmental function is ‘integral’ or ‘traditional’” is “‘unsound in principle and unworkable in practice.’” Indeed, the Court has twice experimented with such standards—first in the context of intergovernmental tax immunity, and more recently in the context of state regulatory immunity under the Commerce Clause—only to abandon them later as analytically unsound. Thus, to the extent today’s holding rests on a distinction between “traditional” governmental functions and their nontraditional counterparts, it cannot be reconciled with prior precedent.

Second, although many municipalities in this country have long assumed responsibility for disposing of local garbage, most of the garbage produced in this country is still managed by the private sector. In that respect, the Court is simply mistaken in concluding that waste disposal is “typically” a local government function.

Moreover, especially considering the Court’s recognition that “‘any notion of discrimination assumes a comparison of substantially similar entities,’” a “traditional” municipal landfill is for present purposes entirely different from a monopolistic landfill supported by the kind of discriminatory legislation at issue in this case and in Carbone. While the former may be rooted in history and tradition, the latter has been deemed unconstitutional until today. It is therefore far from clear that the laws at issue here can fairly be described as serving a function “typically and traditionally” performed by local governments.

C

Equally unpersuasive is the Court’s suggestion that the flow-control laws do not discriminate against interstate commerce because they “treat in-state private business interests exactly the same as out-of-state ones.” Again, the critical issue is whether the challenged legislation discriminates against interstate commerce. If it does, then regardless of whether those harmed by it reside entirely outside the State in question, the law is subject to strict scrutiny. Indeed, this Court has long recognized that “‘a burden imposed by a State upon interstate commerce is not to be sustained simply because the statute imposing it applies alike to the people of all the States, including the people of the State enacting such statute.’” It therefore makes no difference that the flow-control laws at issue here apply to in-state and out-of-state businesses alike. . . .
§ 8. Vouchers and Consumer Choice


The general trend in our times toward increasing intervention by the state in economic affairs has led to a concentration of attention and dispute on the areas where new intervention is proposed and to an acceptance of whatever intervention has so far occurred as natural and unchangeable. The current pause, perhaps reversal, in the trend toward collectivism offers an opportunity to reexamine the existing activities of government and to make a fresh assessment of the activities that are and those that are not justified. This paper attempts such a re-examination for education.

Education is today largely paid for and almost entirely administered by governmental bodies or non-profit institutions. This situation has developed gradually and is now taken so much for granted that little explicit attention is any longer directed to the reasons for the special treatment of education even in countries that are predominantly free enterprise in organization and philosophy. The result has been an indiscriminate extension of governmental responsibility.

The role assigned to government in any particular field depends, of course, on the principles accepted for the organization of society in general. In what follows, I shall assume a society that takes freedom of the individual, or more realistically the family, as its ultimate objective, and seeks to further this objective by relying primarily on voluntary exchange among individuals for the organization of economic activity. In such a free private enterprise exchange economy, government’s primary role is to preserve the rules of the game by enforcing contracts, preventing coercion, and keeping markets free. Beyond this, there are only three major grounds on which government intervention is to be justified. One is “natural monopoly” or similar market imperfection which makes effective competition (and therefore thoroughly voluntary exchange) impossible. A second is the existence of substantial “neighborhood effects,” i.e., the action of one individual imposes significant costs on other individuals for which it is not feasible to make him compensate them or yields significant gains to them for which it is not feasible to make them compensate him—circumstances that again make voluntary exchange impossible. The third derives from an ambiguity in the ultimate objective rather than from the difficulty of achieving it by voluntary exchange, namely, paternalistic concern for children and other irresponsible individuals. The belief in freedom is for “responsible” units, among whom we include neither children nor insane people. In general, this problem is avoided by regarding the family as the basic unit and therefore parents as responsible for their children; in considerable measure, however, such a procedure rests on expediency rather than principle. The problem of drawing a reasonable line between action justified on these paternalistic grounds and action that conflicts with
the freedom of responsible individuals is clearly one to which no satisfactory answer can be given.

In applying these general principles to education, we shall find it helpful to deal separately with (1) general education for citizenship, and (2) specialized vocational education, although it may be difficult to draw a sharp line between them in practice. The grounds for government intervention are widely different in these two areas and justify very different types of action.

**GENERAL EDUCATION FOR CITIZENSHIP**

A stable and democratic society is impossible without widespread acceptance of some common set of values and without a minimum degree of literacy and knowledge on the part of most citizens. Education contributes to both. In consequence, the gain from the education of a child accrues not only to the child or to his parents but to other members of the society; the education of my child contributes to other people’s welfare by promoting a stable and democratic society. Yet it is not feasible to identify the particular individuals (or families) benefited or the money value of the benefit and so to charge for the services rendered. There is therefore a significant “neighborhood effect.”

What kind of governmental action is justified by this particular neighborhood effect? The most obvious is to require that each child receive a minimum amount of education of a specified kind. Such a requirement could be imposed upon the parents without further government action, just as owners of buildings, and frequently of automobiles, are required to adhere to specified standards to protect the safety of others. There is, however, a difference between the two cases. In the latter, individuals who cannot pay the costs of meeting the required standards can generally divest themselves of the property in question by selling it to others who can, so the requirement can readily be enforced without government subsidy—though even here, if the cost of making the property safe exceeds its market value, and the owner is without resources, the government may be driven to paying for the demolition of a dangerous building or the disposal of an abandoned automobile. The separation of a child from a parent who cannot pay for the minimum required education is clearly inconsistent with our reliance on the family as the basic social unit and our belief in the freedom of the individual.

Yet, even so, if the financial burden imposed by such an educational requirement could readily be met by the great bulk of the families in a community, it might be both feasible and desirable to require the parents to meet the cost directly. Extreme cases could be handled by special provisions in much the same way as is done now for housing and automobiles. An even closer analogy is provided by present arrangements for children who are mistreated by their parents. The advantage of imposing the costs on the parents is that it would tend to equalize the social and private costs of having children and so promote a better distribution of families by size.

Differences among families in resources and in number of children—both a reason for and a result of the different policy that has been followed—plus the imposition of a standard of education involving very sizable costs have, however, made such a policy
hardly feasible. Instead, government has assumed the financial costs of providing the education. In doing so, it has paid not only for the minimum amount of education required of all but also for additional education at higher levels available to youngsters but not required of them—as for example in State and municipal colleges and universities. Both steps can be justified by the “neighborhood effect” discussed above—the payment of the costs as the only feasible means of enforcing the required minimum; and the financing of additional education, on the grounds that other people benefit from the education of those of greater ability and interest since this is a way of providing better social and political leadership.

Government subsidy of only certain kinds of education can be justified on these grounds. To anticipate, they do not justify subsidizing purely vocational education which increases the economic productivity of the student but does not train him for either citizenship or leadership. It is clearly extremely difficult to draw a sharp line between these two types of education. Most general education adds to the economic value of the student—indeed it is only in modern times and in a few countries that literacy has ceased to have a marketable value. And much vocational education broadens the student’s outlook. Yet it is equally clear that the distinction is a meaningful one. For example, subsidizing the training of veterinarians, beauticians, dentists, and a host of other specialized skills—as is widely done in the United States in governmentally supported educational institutions—cannot be justified on the same grounds as subsidizing elementary education or, at a higher level, liberal education. Whether it can be justified on quite different grounds is a question that will be discussed later in this paper.

The qualitative argument from the “neighborhood effect” does not, of course, determine the specific kinds of education that should be subsidized or by how much they should be subsidized. The social gain from education is presumably greatest for the very lowest levels of education, where there is the nearest approach to unanimity about the content of the education, and declines continuously as the level of education rises. But even this statement cannot be taken completely for granted—many governments subsidized universities long before they subsidized lower education. What forms of education have the greatest social advantage and how much of the community’s limited resources should be spent on them are questions to be decided by the judgment of the community expressed through its accepted political channels. The role of an economist is not to decide these questions for the community but rather to clarify the issues to be judged by the community in making a choice, in particular, whether the choice is one that it is appropriate or necessary to make on a communal rather than individual basis.

We have seen that both the imposition of a minimum required level of education and the financing of education by the state can be justified by the “neighborhood effects” of education. It is more difficult to justify in these terms a third step that has generally been taken, namely, the actual administration of educational institutions by the government, the “nationalization,” as it were, of the bulk of the “education industry.” The desirability of such nationalization has seldom been faced explicitly because governments have in the main financed education by paying directly the costs of running educational institutions, so that this step has seemed required by the decision to subsidize education. Yet the two steps could readily be separated. Governments could require a minimum level of
education which they could finance by giving parents vouchers redeemable for a specified maximum sum per child per year if spent on “approved” educational services. Parents would then be free to spend this sum and any additional sum on purchasing educational services from an “approved” institution of their own choice. The educational services could be rendered by private enterprises operated for profit, or by non-profit institutions of various kinds. The role of the government would be limited to assuring that the schools met certain minimum standards such as the inclusion of a minimum common content in their programs, much as it now inspects restaurants to assure that they maintain minimum sanitary standards. An excellent example of a program of this sort is the United States educational program for veterans after World War II. Each veteran who qualified was given a maximum sum per year that could be spent at any institution of his choice, provided it met certain minimum standards. A more limited example is the provision in Britain whereby local authorities pay the fees of some students attending non-state schools (the so-called “public schools”). Another is the arrangement in France whereby the state pays part of the costs for students attending non-state schools.

One argument from the “neighborhood effect” for nationalizing education is that it might otherwise be impossible to provide the common core of values deemed requisite for social stability. The imposition of minimum standards on privately conducted schools, as suggested above, might not be enough to achieve this result. The issue can be illustrated concretely in terms of schools run by religious groups. Schools run by different religious groups will, it can be argued, instill sets of values that are inconsistent with one another and with those instilled in other schools; in this way they convert education into a divisive rather than a unifying force.

Carried to its extreme, this argument would call not only for governmentally administered schools, but also for compulsory attendance at such schools. Existing arrangements in the United States and most other Western countries are a halfway house. Governmentally administered schools are available but not required. However, the link between the financing of education and its administration places other schools at a disadvantage: they get the benefit of little or none of the governmental funds spent on education—a situation that has been the source of much political dispute, particularly, of course, in France. The elimination of this disadvantage might, it is feared, greatly strengthen the parochial schools and so render the problem of achieving a common core of values even more difficult.

This argument has considerable force. But it is by no means clear either that it is valid or that the denationalizing of education would have the effects suggested. On grounds of principle, it conflicts with the preservation of freedom itself; indeed, this conflict was a major factor retarding the development of state education in England. How draw a line between providing for the common social values required for a stable society on the one hand, and indoctrination inhibiting freedom of thought and belief on the other? Here is an other of those vague boundaries that it is easier to mention than to define.

In terms of effects, the denationalization of education would widen the range of choice available to parents. Given, as at present, that parents can send their children to government schools without special payment, very few can or will send them to other
schools unless they too are subsidized. Parochial schools are at a disadvantage in not getting any of the public funds devoted to education; but they have the compensating advantage of being run by institutions that are willing to subsidize them and can raise funds to do so, whereas there are few other sources of subsidies for schools. Let the subsidy be made available to parents regardless where they send their children—provided only that it be to schools that satisfy specified minimum standards—and a wide variety of schools will spring up to meet the demand. Parents could express their views about schools directly, by withdrawing their children from one school and sending them to another, to a much greater extent than is now possible. In general, they can now take this step only by simultaneously changing their place of residence. For the rest, they can express their views only through cumbrous political channels. Perhaps a somewhat greater degree of freedom to choose schools could be made available also in a governmentally administered system, but it is hard to see how it could be carried very far in view of the obligation to provide every child with a place. Here, as in other fields, competitive private enterprise is likely to be far more efficient in meeting consumer demands than either nationalized enterprises or enterprises run to serve other purposes. The final result may therefore well be less rather than more parochial education.

Another special case of the argument that governmentally conducted schools are necessary to keep education a unifying force is that private schools would tend to exacerbate class distinctions. Given greater freedom about where to send their children, parents of a kind would flock together and so prevent a healthy intermingling of children from decidedly different backgrounds. Again, whether or not this argument is valid in principle, it is not at all clear that the stated results would follow. Under present arrangements, particular schools tend to be peopled by children with similar backgrounds thanks to the stratification of residential areas. In addition, parents are not now prevented from sending their children to private schools. Only a highly limited class can or does do so, parochial schools aside, in the process producing further stratification. The widening of the range of choice under a private system would operate to reduce both kinds of stratification.

Another argument for nationalizing education is “natural monopoly.” In small communities and rural areas, the number of children may be too small to justify more than one school of reasonable size, so that competition cannot be relied on to protect the interests of parents and children. As in other cases of natural monopoly, the alternatives are unrestricted private monopoly, state-controlled private monopoly, and public operation—a choice among evils. This argument is clearly valid and significant, although its force has been greatly weakened in recent decades by improvements in transportation and increasing concentration of the population in urban communities.

The arrangement that perhaps comes closest to being justified by these considerations—at least for primary and secondary education—is a mixed one under which governments would continue to administer some schools but parents who chose to send their children to other schools would be paid a sum equal to the estimated cost of educating a child in a government school, provided that at least this sum was spent on education in an approved school. This arrangement would meet the valid features of the “natural monopoly” argument, while at the same time it would permit competition to develop where it could.
It would meet the just complaints of parents that if they send their children to private nonsubsidized schools they are required to pay twice for education—once in the form of general taxes and once directly—and in this way stimulate the development and improvement of such schools. The interjection of competition would do much to promote a healthy variety of schools. It would do much, also, to introduce flexibility into school systems. Not least of its benefits would be to make the salaries of school teachers responsive to market forces. It would thereby give governmental educational authorities an independent standard against which to judge salary scales and promote a more rapid adjustment to changes in conditions of demand or supply.  

Why is it that our educational system has not developed along these lines? A full answer would require a much more detailed knowledge of educational history than I possess, and

---

2 Essentially this proposal—public financing but private operation of education—has recently been suggested in several southern states as a means of evading the Supreme Court ruling against segregation. This fact came to my attention after this paper was essentially in its present form. My initial reaction—and I venture to predict, that of most readers—was that this possible use of the proposal was a count against it, that it was a particularly striking case of the possible defect—the exacerbating of class distinctions—referred to in the second paragraph preceding the one to which this note is attached.

Further thought has led me to reverse my initial reaction. Principles can be tested most clearly by extreme cases. Willingness to permit free speech to people with whom one agrees is hardly evidence of devotion to the principle of free speech; the relevant test is willingness to permit free speech to people with whom one thoroughly disagrees. Similarly, the relevant test of the belief in individual freedom is the willingness to oppose state intervention even when it is designed to prevent individual activity of a kind one thoroughly dislikes. I deplore segregation and racial prejudice; pursuant to the principles set forth at the outset of the paper, it is clearly an appropriate function of the state to prevent the use of violence and physical coercion by one group on another; equally clearly, it is not an appropriate function of the state to try to force individuals to act in accordance with my—or anyone else’s—views, whether about racial prejudice or the party to vote for; so long as the action of any one individual affects mostly himself. These are the grounds on which I oppose the proposed Fair Employment Practices Commissions; and they lead me equally to oppose forced nonsegregation. However, the same grounds also lead me to oppose forced segregation. Yet, so long as the schools are publicly operated, the only choice is between forced nonsegregation and forced segregation; and if I must choose between these evils, I would choose the former as the lesser. The fact that I must make this choice is a reflection of the basic weakness of a publicly operated school system. Privately conducted schools can resolve the dilemma. They make unnecessary either choice. Under such a system, there can develop exclusively white schools, exclusively colored schools, and mixed schools. Parents can choose which to send their children to. The appropriate activity for those who oppose segregation and racial prejudice is to try to persuade others of their views; and as they succeed, the mixed schools will grow at the expense of the nonmixed, and a gradual transition will take place. So long as the school system is publicly operated, only drastic change is possible; one must go from one extreme to the other; it is a great virtue of the private arrangement that it permits a gradual transition.

An example that comes to mind as illustrating the preceding argument is summer camps for children. Is there any objection to the simultaneous existence of some camps that are wholly Jewish, some wholly non-Jewish, and some mixed? One can—though many who would react quite differently to negro-white segregation would not—deplore the existence of attitudes that lead to the three types: one can seek to propagate views that would tend to the growth of the mixed school at the expense of the extremes; but is it an appropriate function of the state to prohibit the unmixed camps?

The establishment of private schools does not of itself guarantee the desirable freedom of choice on the part of parents. The public funds could be made available subject to the condition that parents use them solely in segregated schools; and it may be that some such condition is contained in the proposals now under consideration by southern states. Similarly, the public funds could be made available for use solely in nonsegregated schools. The proposed plan is not therefore inconsistent with forced segregation or forced nonsegregation. The point is that it makes available a third alternative.
the most I can do is to offer a conjecture. For one thing, the “natural monopoly” argument was much stronger at an earlier date. But I suspect that a much more important factor was the combination of the general disrepute of cash grants to individuals (“handouts”) with the absence of an efficient administrative machinery to handle the distribution of vouchers and to check their use. The development of such machinery is a phenomenon of modern times that has come to full flower only with the enormous extension of personal taxation and of social security programs. In its absence, the administration of schools was regarded as the only possible way to finance education. Of course, as some of the examples cited above suggest, some features of the proposed arrangements are present in existing educational systems. And there has been strong and I believe increasing pressure for arrangements of this general kind in most Western countries, which is perhaps to be explained by the modern developments in governmental administrative machinery that facilitate such arrangements.

Many detailed administrative problems would arise in changing over from the present to the proposed system and in administering the proposed system. But these seem neither insoluble nor unique. As in the denationalization of other activities, existing premises and equipment could be sold to private enterprises that wanted to enter the field, so there would be no waste of capital in the transition. The fact that governmental units, at least in some areas, were going to continue to administer schools would permit a gradual and easy transition. The localized administration of education in the United States and some other countries would similarly facilitate the transition, since it would encourage experimentation on a small scale and with alternative methods of handling both these and other problems. Difficulties would doubtless arise in determining eligibility for grants from a particular governmental unit, but this is identical with the existing problem of determining which unit is obligated to provide educational facilities for a particular child. Differences in size of grants would make one area more attractive than another just as differences in the quality of education now have the same effect. The only additional complication is a possibly greater opportunity for abuse because of the greater freedom to decide where to educate children. Supposed difficulty of administration is a standard defense of the status quo against any proposed changes; in this particular case, it is an even weaker defense than usual because existing arrangements must master not only the major problems raised by the proposed arrangements but also the additional problems raised by the administration of the schools as a governmental function.

The preceding discussion is concerned mostly with primary and secondary education. For higher education, the case for nationalization on grounds either of neighborhood effects or of natural monopoly is even weaker than for primary and secondary education. For the lowest levels of education, there is considerable agreement, approximating unanimity, on the appropriate content of an educational program for citizens of a democracy—the three R’s cover most of the ground. At successively higher levels of education, there is less and less agreement. Surely, well below the level of the American college, one can expect insufficient agreement to justify imposing the views of a majority, much less a plurality, on all. The lack of agreement may, indeed, extend so far as to cast doubts on the appropriateness of even subsidizing education at this level; it surely goes far enough to undermine any case for nationalization on the grounds of providing a common core of values. Similarly, there can hardly be any question of “natural monopoly” at this level, in
view of the distances that individuals can and do go to attend institutions of higher learning.

Governmental institutions in fact play a smaller role in the United States in higher education than at lower levels. Yet they grew greatly in importance until at least the 1920’s and now account for more than half the students attending colleges and universities. One of the main reasons for their growth was their relative cheapness: most State and municipal colleges and universities charge much lower tuition fees than private universities can afford to. Private universities have in consequence had serious financial problems, and have quite properly complained of “unfair” competition. They have wanted to maintain their independence from government, yet at the same time have felt driven by financial pressure to seek government aid.

The preceding analysis suggests the lines along which a satisfactory solution can be found. Public expenditure on higher education can be justified as a means of training youngsters for citizenship and for community leadership—though I hasten to add that the large fraction of current expenditure that goes for strictly vocational training cannot be justified in this way or, indeed, as we shall see, in any other. Restricting the subsidy to education obtained at a state-administered institution cannot be justified on these grounds, or on any other that I can derive from the basic principles outlined at the outset. Any subsidy should be granted to individuals to be spent at institutions of their own choosing, provided only that the education is of a kind that it is desired to subsidize. Any government schools that are retained should charge fees covering the cost of educating students and so compete on an equal level with non-government-supported schools. The retention of state schools themselves would, however, have to be justified on grounds other than those we have so far considered. The resulting system would follow in its broad outlines the arrangements adopted in the United States after World War II for financing the education of veterans, except that the funds would presumably come from the States rather than the Federal government.

The adoption of such arrangements would make for more effective competition among various types of schools and for a more efficient utilization of their resources. It would eliminate the pressure for direct government assistance to private colleges and universities and thus preserve their full independence and diversity at the same time that it enabled them to grow relatively to State institutions. It might also have the ancillary advantage of causing a closer scrutiny of the purposes for which subsidies are granted. The subsidization of institutions rather than of people has led to an indiscriminate subsidization of whatever activities it is appropriate for such institutions to undertake, rather than of the activities it is appropriate for the state to subsidize. Even cursory examination suggests that while the two classes of activities overlap, they are far from identical.

**Vocational or Professional Education**

As noted above, vocational or professional education has no neighborhood effects of the kind attributed above to general education. It is a form of investment in human capital precisely analogous to investment in machinery, buildings, or other forms of non human
capital. Its function is to raise the economic productivity of the human being. If it does so, the individual is rewarded in a free enterprise society by receiving a higher return for his services than he would otherwise be able to command. This difference is the economic incentive to acquire the specialized training, just as the extra return that can be obtained with an extra machine is the economic incentive to invest capital in the machine. In both cases, extra returns must be balanced against the costs of acquiring them. For vocational education, the major costs are the income foregone during the period of training, interest lost by postponing the beginning of the earning period, and special expenses of acquiring the training such as tuition fees and expenditures on books and equipment. For physical capital, the major costs are the expenses of constructing the capital equipment and the interest during construction. In both cases, an individual presumably regards the investment as desirable if the extra returns, as he views them, exceed the extra costs, as he views them. In both cases, if the individual undertakes the investment and if the state neither subsidizes the investment nor taxes the return, the individual (or his parent, sponsor, or benefactor) in general bears all the extra cost and receives all the extra returns: there are no obvious unborne costs or unappropriable returns that tend to make private incentives diverge systematically from those that are socially appropriate.

If capital were as readily available for investment in human beings as for investment in physical assets, whether through the market or through direct investment by the individuals concerned or their parents or benefactors, the rate of return on capital would tend to be roughly equal in the two fields: if it were higher on non-human capital, parents would have an incentive to buy such capital for their children instead of investing a corresponding sum in vocational training, and conversely. In fact, however, there is considerable empirical evidence that the rate of return on investment in training is very much higher than the rate of return on investment in physical capital. According to estimates that Simon Kuznets and I have made elsewhere, professionally trained workers in the United States would have had to earn during the 1930’s almost 70 per cent more than other workers to cover the extra costs of their training, including interest at roughly the market rate on non-human capital. In fact, they earned on the average between two and three times as much. Some part of this difference may well be attributable to greater natural ability on the part of those who entered the professions: it may be that they would have earned more than the average non-professional worker if they had not gone into the professions. Kuznets and I concluded, however, that such differences in ability could not explain anything like the whole of the extra return of the professional workers. Apparently, there was sizable underinvestment in human beings. The postwar period has doubtless brought changes in the relative earnings in different occupations. It seems extremely doubtful, however, that they have been sufficiently great to reverse this conclusion.

It is not certain at what level this underinvestment sets in. It clearly applies to professions requiring a long period of training, such as medicine, law, dentistry, and the like, and probably to all occupations requiring a college training. At one time, it almost certainly extended to many occupations requiring much less training but probably no longer does, although the opposite has some times been maintained.
This underinvestment in human capital presumably reflects an imperfection in the capital market: investment in human beings cannot be financed on the same terms or with the same ease as investment in physical capital. It is easy to see why there would be such a difference. If a fixed money loan is made to finance investment in physical capital, the lender can get some security for his loan in the form of a mortgage or residual claim to the physical asset itself, and he can count on realizing at least part of his investment in case of necessity by selling the physical asset. If he makes a comparable loan to increase the earning power of a human being, he clearly cannot get any comparable security; in a non-slave state, the individual embodying the investment cannot be bought and sold. But even if he could, the security would not be comparable. The productivity of the physical capital does not—or at least generally does not—depend on the co-operativeness of the original borrower. The productivity of the human capital quite obviously does—which is, of course, why, all ethical considerations aside, slavery is economically inefficient. A loan to finance the training of an individual who has no security to offer other than his future earnings is therefore a much less attractive proposition than a loan to finance, say, the erection of a building: the security is less, and the cost of subsequent collection of interest and principal is very much greater.

A further complication is introduced by the inappropriateness of fixed money loans to finance investment in training. Such an investment necessarily involves much risk. The average expected return may be high, but there is wide variation about the average. Death or physical incapacity is one obvious source of variation but is probably much less important than differences in ability, energy, and good fortune. The result is that if fixed money loans were made, and were secured only by expected future earnings, a considerable fraction would never be repaid. In order to make such loans attractive to lenders, the nominal interest rate charged on all loans would have to be sufficiently high to compensate for the capital losses on the defaulted loans. The high nominal interest rate would both conflict with usury laws and make the loans unattractive to borrowers, especially to borrowers who have or expect to have other assets on which they cannot currently borrow but which they might have to realize or dispose of to pay the interest and principal of the loan. The device adopted to meet the corresponding problem for other risky investments is equity investment plus limited liability on the part of shareholders. The counterpart for education would be to “buy” a share in an individual’s earning prospects: to advance him the funds needed to finance his training on condition that he agree to pay the lender a specified fraction of his future earnings. In this way, a lender would get back more than his initial investment from relatively successful individuals, which would compensate for the failure to recoup his original investment from the unsuccessful.

There seems no legal obstacle to private contracts of this kind, even though they are economically equivalent to the purchase of a share in an individual’s earning capacity and thus to partial slavery. One reason why such contracts have not become common, despite their potential profitability to both lenders and borrowers, is presumably the high costs of administering them, given the freedom of individuals to move from one place to another, the need for getting accurate income statements, and the long period over which the contracts would run. These costs would presumably be particularly high for investment on a small scale with a resultant wide geographical spread of the individuals.
financed in this way. Such costs may well be the primary reason why this type of investment has never developed under private auspices. But I have never been able to persuade myself that a major role has not also been played by the cumulative effect of such factors as the novelty of the idea, the reluctance to think of investment in human beings as strictly comparable to investment in physical assets, the resultant likelihood of irrational public condemnation of such contracts, even if voluntarily entered into, and legal and conventional limitation on the kind of investments that may be made by the financial intermediaries that would be best suited to engage in such investments, namely, life insurance companies. The potential gains, particularly to early entrants, are so great that it would be worth incurring extremely heavy administrative costs.

But whatever the reason, there is clearly here an imperfection of the market that has led to underinvestment in human capital and that justifies government intervention on grounds both of “natural monopoly,” insofar as the obstacle to the development of such investment has been administrative costs, and of improving the operation of the market, insofar as it has been simply market frictions and rigidities.

What form should government intervention take? One obvious form, and the only form that it has so far taken, is outright government subsidy of vocational or professional education financed out of general revenues. Yet this form seems clearly inappropriate. Investment should be carried to the point at which the extra return repays the investment and yields the market rate of interest on it. If the investment is in a human being, the extra return takes the form of a higher payment for the individual’s services than he could otherwise command. In a private market economy, the individual would get this return as his personal income, yet if the investment were subsidized, he would have borne none of the costs. In consequence, if subsidies were given to all who wished to get the training, and could meet minimum quality standards, there would tend to be overinvestment in human beings, for individuals would have an incentive to get the training so long as it yielded any extra return over private costs, even if the return were insufficient to repay the capital invested, let alone yield any interest on it. To avoid such overinvestment, government would have to restrict the subsidies. Even apart from the difficulty of calculating the “correct” amount of investment, this would involve rationing in some essentially arbitrary way the limited amount of investment among more claimants than could be financed, and would mean that those fortunate enough to get their training subsidized would receive all the returns from the investment whereas the costs would be borne by the taxpayers in general. This seems an entirely arbitrary, if not perverse, redistribution of income.

The desideratum is not to redistribute income but to make capital available for investment in human beings on terms comparable to those on which it is available for physical investment. Individuals should bear the costs of investment in themselves and receive the rewards, and they should not be prevented by market imperfections from making the investment when they are willing to bear the costs. One way to do this is to have governmental body engage in equity investment in human beings of the kind described above. A governmental body could offer to finance or help finance the training of any individual who could meet minimum quality standards by making available not more than a limited sum per year for not more than a specified number of years, provided it was spent on
securing training at a recognized institution. The individual would agree in return to pay to the government in each future year $x$ per cent of his earnings in excess of $y$ dollars for each $1,000 that he gets in this way. This payment could easily be combined with payment of income tax and so involve a minimum of additional administrative expense. The base sum, $y$, should be set equal to estimated average—or perhaps modal—earnings without the specialized training; the fraction of earnings paid, $x$, should be calculated so as to make the whole project self-financing. In this way the individuals who received the training would in effect bear the whole cost. The amount invested could then be left to be determined by individual choice. Provided this was the only way in which government financed vocational or professional training, and provided the calculated earnings reflected all relevant returns and costs, the free choice of individuals would tend to produce the optimum amount of investment.

The second proviso is unfortunately not likely to be fully satisfied. In practice, therefore, investment under the plan would still be somewhat too small and would not be distributed in the optimum manner. To illustrate the point at issue, suppose that a particular skill acquired by education can be used in two different ways; for example, medical skill in research or in private practice. Suppose that, if money earnings were the same, individuals would generally prefer research. The non-pecuniary advantages of research would then tend to be offset by higher money earnings in private practice. These higher earnings would be included in the sum to which the fraction $x$ was applied whereas the monetary equivalent of the non-pecuniary advantages of research would not be. In consequence, the earnings differential would have to be higher under the plan than if individuals could finance themselves, since it is the net monetary differential, not the gross, that individuals would balance against the non-pecuniary advantages of research in deciding how to use their skill. This result would be produced by a larger than optimum fraction of individuals going into research necessitating a higher value of $x$ to make the scheme self-financing than if the value of the non-pecuniary advantages could be included in calculated earnings. The inappropriate use of human capital financed under the plan would in this way lead to a less than optimum incentive to invest and so to a less than optimum amount of investment.

Estimation of the values of $x$ and $y$ clearly offers considerable difficulties, especially in the early years of operation of the plan, and the danger would always be present that they would become political footballs. Information on existing earnings in various occupations is relevant but would hardly permit anything more than a rough approximation to the values that would render the project self-financing. In addition, the values should in principle vary from individual to individual in accordance with any differences in expected earning capacity that can be predicted in advance—the problem is similar to that of varying life insurance premia among groups that have different life expectancy. For such reasons as these it would be preferable if similar arrangements could be developed on a private basis by financial institutions in search of outlets for investing their funds, non-profit institutions such as private foundations, or individual universities and colleges.

Insofar as administrative expense is the obstacle to the development of such arrangements on a private basis, the appropriate unit of government to make funds available is the
Federal government in the United States rather than smaller units. Any one State would have the same costs as an insurance company, say, in keeping track of the people whom it had financed. These would be minimized for the Federal government. Even so, they would not be completely eliminated. An individual who migrated to another country, for example, might still be legally or morally obligated to pay the agreed-on share of his earnings, yet it might be difficult and expensive to enforce the obligation. Highly successful people might therefore have an incentive to migrate. A similar problem arises, of course, also under the income tax, and to a very much greater extent. This and other administrative problems of conducting the scheme on a Federal level, while doubtless troublesome in detail, do not seem serious. The really serious problem is the political one already mentioned: how to prevent the scheme from becoming a political football and in the process being converted from a self-financing project to a means of subsidizing vocational education.

But if the danger is real, so is the opportunity. Existing imperfections in the capital market tend to restrict the more expensive vocational and professional training to individuals whose parents or benefactors can finance the training required. They make such individuals a “non-competing” group sheltered from competition by the unavailability of the necessary capital to many individuals, among whom must be large numbers with equal ability. The result is to perpetuate inequalities in wealth and status. The development of arrangements such as those outlined above would make capital more widely available and would thereby do much to make equality of opportunity a reality, to diminish inequalities of income and wealth, and to promote the full use of our human resources. And it would do so not, like the outright redistribution of income, by impeding competition, destroying incentive, and dealing with symptoms, but by strengthening competition, making incentives effective, and eliminating the causes of inequality.

CONCLUSION

This re-examination of the role of government in education suggests that the growth of governmental responsibility in this area has been unbalanced. Government has appropriately financed general education for citizenship, but in the process it has been led also to administer most of the schools that provide such education. Yet, as we have seen, the administration of schools is neither required by the financing of education, nor justifiable in its own right in a predominantly free enterprise society. Government has appropriately been concerned with widening the opportunity of young men and women to get professional and technical training, but it has sought to further this objective by the inappropriate means of subsidizing such education, largely in the form of making it available free or at a low price at governmentally operated schools.

The lack of balance in governmental activity reflects primarily the failure to separate sharply the question what activities it is appropriate for government to finance from the question what activities it is appropriate for government to administer—a distinction that is important in other areas of government activity as well. Because the financing of general education by government is widely accepted, the provision of general education directly by governmental bodies has also been accepted. But institutions that provide general education are especially well suited also to provide some kinds of vocational and
professional education, so the acceptance of direct government provision of general education has led to the direct provision of vocational education. To complete the circle, the provision of vocational education has, in turn, meant that it too was financed by government, since financing has been predominantly of educational institutions not of particular kinds of educational services.

The alternative arrangements whose broad outlines are sketched in this paper distinguish sharply between the financing of education and the operation of educational institutions, and between education for citizenship or leadership and for greater economic productivity. Throughout, they center attention on the person rather than the institution. Government, preferably local governmental units, would give each child, through his parents, a specified sum to be used solely in paying for his general education; the parents would be free to spend this sum at a school of their own choice, provided it met certain minimum standards laid down by the appropriate governmental unit. Such schools would be conducted under a variety of auspices: by private enterprises operated for profit, non-profit institutions established by private endowment, religious bodies, and some even by governmental units.

For vocational education, the government, this time however the central government, might likewise deal directly with the individual seeking such education. If it did so, it would make funds available to him to finance his education, not as a subsidy but as “equity” capital. In return, he would obligate himself to pay the state a specified fraction of his earnings above some minimum, the fraction and minimum being determined to make the program self-financing. Such a program would eliminate existing imperfections in the capital market and so widen the opportunity of individuals to make productive investments in themselves while at the same time assuring that the costs are borne by those who benefit most directly rather than by the population at large. An alternative, and a highly desirable one if it is feasible, is to stimulate private arrangements directed toward the same end.

The result of these measures would be a sizable reduction in the direct activities of government, yet a great widening in the educational opportunities open to our children. They would bring a healthy increase in the variety of educational institutions available and in competition among them. Private initiative and enterprise would quicken the pace of progress in this area as it has in so many others. Government would serve its proper function of improving the operation of the invisible hand without substituting the dead hand of bureaucracy.
§ 8.2. School Vouchers


SUMMARY

The most frequently asked questions about school choice are: Do public schools respond constructively to competition induced by school choice, by raising their own productivity? Does students’ achievement rise when they attend voucher or charter schools? Do voucher and charter schools end up with a selection of the better students (“cream-skim”)? I review the evidence on these questions from the United States, relying primarily on recent policy experiments. Public schools do respond constructively to competition, by raising their achievement and productivity. The best studies on this question examine the introduction of choice programs that have been sufficiently large and long-lived to produce competition. Students’ achievement generally does rise when they attend voucher or charter schools. The best studies on this question use, as a control group, students who are randomized out of choice programs. Not only do currently enacted voucher and charter school programs not cream-skim; they disproportionately attract students who were performing badly in their regular public schools. This confirms what theory predicts: there are no general results on the sorting consequences of school choice. The sorting consequences of a school choice plan depend strongly on its design.

The first generation of noteworthy school choice programs in the United States were enacted in the period from 1988 to 1994. Approximately a decade later, it is appropriate to assess what we have learned from them. An additional motivation for such an assessment is that school choice legislation is on the verge of a second wave of activity, owing to the June 2002 decision of the United States Supreme Court (Zelman v. Simmons-Harris). The Supreme Court upheld the constitutionality of a voucher program in Cleveland, Ohio that included private schools with religious affiliation; and the Court described which future programs would be constitutional. To be constitutional, school choice programs must not be designed in such a way that a student who preferred to choose a secular school was induced to choose a religious school. The school choice plans that are currently under contemplation already meet this condition. As a result, the Supreme Court decision has opened the door for several years worth of pent-up legislation. Momentum for school choice in central cities has also gained strength from the recent movement toward statewide testing (codified by the federal No Child Left Behind Act of 2002). School report cards have revealed to the general public what researchers have known for some time: an extraordinary prevalence of failure at many central city schools, despite their having greater per-pupil revenues than most other
schools. Such revelations are fueling campaigns for school choice in areas like the District of Columbia.

Because second-generation school choice programs are currently in an active planning stage, it makes sense to assemble what we have learned so far. This is the agenda for this paper. I attempt to answer the three most common questions about school choice, relying primarily on evidence from first generation programs in the United States. The questions are: Do public schools respond constructively to competition induced by school choice, by raising their own productivity? Does students’ achievement rise when they attend voucher or charter schools? Do voucher and charter schools end up with a selection of the better students (“cream-skim”)? Throughout, I focus on evidence that is recent and that relies on the most credible empirical methods. Because the effects of school choice programs depend on their design and the public school environment on which they are superimposed, I do not attempt to review international evidence: accurate descriptions would necessarily occupy too much space.

1. FORMS OF CHOICE IN THE US: VOUCHERS AND CHARTER SCHOOLS

In the US, school choice takes two main forms: vouchers and charter schools.

1.1. Vouchers

A voucher is a coupon that a student carries with him to the school of his choice, at least some of which are potentially private. When he enrolls, the school gets revenue equal to the amount of his voucher. Public money funds the voucher. (Privately funded vouchers currently exist, but they are intended as experiments, to guide future policies. The policy version of vouchers is always intended to be publicly funded.)

I have now said everything is that is true of all voucher policies. This is because a voucher is inherently a flexible instrument that can be designed in many ways. For instance, a voucher may be designed for use at public and private schools, at all private schools, or only at private schools that meet certain criteria (for instance, secular or accredited private schools). Voucher-accepting schools are usually required to admit voucher applicants by lottery but may be allowed to practice selective admission. The public money that funds the voucher may come from local, state, or federal budgets. “Topping up” the voucher may or may not be permitted (topping up occurs when a school is allowed to charge students tuition in excess of the voucher, thus requiring parents to “top up” from their own income). The amount of a voucher can be identical across all students, can vary with a student’s characteristics, or can vary with the characteristics of both a student and his school. In theory, the voucher amount can range widely. In practice, the typical voucher in the US is worth between 14 and 29 percent of per-pupil expenditure in the local public schools. . . . One of the most generous voucher programs in the US (Milwaukee) offers vouchers equal to about half of local per-pupil spending. In the 2002-03 school year, the Milwaukee voucher was worth USD 5,783 and the Milwaukee Public Schools spent USD 11,436 per pupil. Currently, there is no voucher equal to local per-pupil expenditure.
Hereafter, I sometimes use the phrase “voucher school” to describe a private school that enrolls students with vouchers.

1.2. Charter Schools

Charter schools are schools chartered by a government or government-appointed body to educate children in return for a publicly funded fee (the “charter school fee”). It may be best to think of charter policies as voucher policies with more constraints imposed on the schools. Although the constraints on charter schools vary widely among states, certain constraints are always imposed. First, charter schools are never allowed to practice “positive” selective admissions in the sense of excluding students with poor test scores or interviews. They are usually required to accept students by lottery or on the basis of certain characteristics that are considered “negative” (for instance, a student’s having indicated that he would like to drop out of school). Second, charter schools must accept the charter school fee; they cannot allow or require parents to top up the fee. Third, charter schools are legally public institutions, so they must obey the same regulations on church-state relations, racial and gender discrimination, et cetera as public schools. Fourth, the chartering process requires the school to meet certain government designated criteria at regular intervals: at the initiation of the charter school and again at periodic rechartering. The chartering criteria may be more or less restrictive—in a few states, they are almost indistinguishable from the accreditation criteria for private schools. In other states, the criteria are more stringent than the accreditation criteria for public schools. Because the chartering body is a government or is government appointed, all charter schools are somewhat vulnerable to political attacks.

In short, charter schools are always distinguishable from vouchers, but, on the school choice spectrum, the most restrictive voucher programs run pretty smoothly into the least restrictive charter school programs. The average charter school in the US gets a per-pupil fee equal to 45 percent of what its local public school competitors spend.

2. Does Competition Force Public Schools to Be More Productive?

A school that is more productive is one that produces higher achievement in its pupils for each dollar it spends. Formally, a school’s productivity is defined as achievement per dollar spent, controlling for incoming achievement differences of its students. In practice, although it makes the most sense to ask whether competition induces public schools to be more productive, we usually separately check the effect on achievement. This is because we may not feel symmetrically about productivity gains that are attained through achievement gains (keeping current costs roughly equal) and productivity gains attained through cost cuts (keeping current achievement roughly equal).

The question of whether school choice induces regular public schools to be more productive is an important one. This is for two reasons. First, in the short term, public schools’ exhibiting a positive productivity response would greatly extend the benefits of school choice beyond the students who are first to take up the opportunity to attend voucher or charter schools. Second, when advocates of school choice argue that every child would benefit from school choice, they are usually relying on the idea that school
productivity would increase sufficiently to swamp any negative allocation effects that some students might experience. In other words, a general increase in school productivity could be a rising tide that lifted all boats, and the gains and losses from reallocation might be nothing more than crests and valleys on the surface of the much higher water level. (An allocation effect is any effect resulting from the reallocation of students among schools. Such effects might operate through peer effects, house prices, or politics. See below for more on this issue.)

The hypothesis about choice and productivity is quite straightforward: when students can leave and money follows students (even if imperfectly or indirectly), less productive schools will lose students to more productive schools. That is, if school X . . . could raise a student’s achievement while spending the same amount as his current school, then school X would be expected to draw him away from his current school. This process would shrink the less productive and expand the more productive schools, until one of two things happened: the more productive replaced the less productive school or the less productive school raised its productivity and was thereby able to maintain its population of students. (This is merely the essential intuition, see Hoxby (2003) for a formal exposition that describes detailed mechanisms for regular public schools, voucher schools, and charter schools.)

Obviously, there are pseudo-choice plans that would not allow anything like a competitive process to occur. These are plans in which money does not follow students or so little money follows students that a school accepting an extra student cannot cover its marginal costs; plans in which schools are not able to enter, expand, contract, or exit; plans in which schools need to seek approval or financial support from other schools with which they are supposed to compete; and so on. Indeed, pseudo-choice plans that lack the semblance of a competitive structure are common; and plans that provide perverse incentives (because they punish successful schools by decreasing their funding and vice versa) are not at all rare. Some empirical researchers have studied such pseudo-choice plans under the heading of school choice and have discovered that the pseudo-choice plans do not, in fact, exhibit conduct we expect from a competitive marketplace. Such research confuses readers unnecessarily. Economists are responsible for describing the structure of the market they are analyzing and testing hypotheses that are based on that structure. The study of educational markets does not free them from this responsibility. (Think of an industrial organization economist not bothering to point out whether a market has a structure that is monopoly, oligopoly, perfect competition, et cetera.) Far too often in the empirical literature on school choice, the reader is presented with results but no clear idea of whether the plan under analysis fulfills even basic criteria for being a school choice plan.

2.1. Is It Reasonable to Think That Public Schools Could Be More Productive?

One does not have to be hopelessly optimistic to think that regular public schools in the US could be substantially more productive. Straightforward productivity estimates suggest that their productivity was approximately 65 percent higher in 1970-71 than in 1998-99, the most recent year for which we can produce estimates. The 65 percent figure is what one gets by dividing National Assessment of Educational Progress scores by per-
pupil spending, adjusted by the Consumer Price Index. If one adjusts for the family background of student test-takers, the estimated decline in productivity grows slightly larger, to 68 or 69 percent. (The 68 percent figure is based on using coefficients from 1998-99 and the 69 percent figure is based on using coefficients from 1970-71). Even if one adjusts for the increased cost of hiring able people as teachers by deflating spending with an index of the wages of women with professional degrees, the estimated decline in productivity is still a substantial 50 percent. (Note that this method of spending deflation addresses Baumol’s “cost disease” argument.) Deflating by professional women’s wages is certainly an over-adjustment for cost increases: teacher compensation makes up only two-thirds percent of a typical school’s budget and the wages of professional women went up significantly faster than those of women who were merely baccalaureate degree holders. American teachers have never been drawn primarily from the pool of women with the ability and skills of professional degree holders. For instance, Corcoran, Evans, and Schwab (2002) show that the typical female teacher of 1957 to 1963 scored 7 percentiles below the average female college graduate, let alone the female who was likely to attain a professional degree.

In short, the evidence suggests that there is room for productivity gains of 50 to 69 percent. We are not accustomed to numbers this large. One could presumably introduce a variety of further adjustments, but the main ones have been covered and there is not much point in tweaking the numbers further: the question at issue is simply whether regular public schools are capable of higher productivity than they currently exhibit. The evidence suggests that they are.

2.2. Empirical Approaches to Estimating the Effect of School Choice and Competition on Productivity

In other work, I have studied the effects of traditional forms of school choice in the US, especially Tiebout choice. Some American metropolitan areas have traditionally enjoyed Tiebout choice to a great extent, owing to the highly local nature of their school districts and school finance. In this study, I will not discuss evidence on traditional forms of choice, simply because it would require too much explanation. It is useful, however, to point out what we lose by the omission: evidence on the long-term, general equilibrium effects of choice.

In the short term, an administrator who is attempting to raise his school’s productivity has only certain options. He can induce his staff to work harder; he can get rid of unproductive staff and programs; he can allocate resources away from non-achievement oriented activities and toward achievement oriented ones. In the slightly longer term, he can renegotiate the teacher contract to make the school more efficient. If an administrator pursues all of these options, he may be able to raise productivity substantially.

Nevertheless, choice can affect productivity through a variety of long-term, general equilibrium mechanisms that are not immediately available to an administrator. The financial pressures of choice may bid up the wages of teachers whose teaching raises achievement and attracts parents. It may thus draw people into teaching (or keep people in teaching) who would otherwise pursue other careers. Indeed, there is evidence that
schools under pressure from choice reward teachers more on the basis of merit and allow administrators more discretion in rewarding good teachers (see Hoxby, 2002, and Hanushek and Rivkin, 2003). The need to attract parents may force schools to issue more information about their achievement and may thus gradually make parents into better “consumers.” Because parents’ decisions are more meaningful when schools are financed by fees they control, choice may make schools more receptive to parent participation. The need to produce results that are competitive with those of other schools may force schools to recognize and abandon pedagogical techniques and curricula that are unsuccessful in practice though philosophically appealing. Finally, in the long-term, choice can affect the size and very existence of schools. Choice makes districts’ enrollment expand and contract; it makes schools enter and exit. In the short term, we mainly observe how the existing stock of schools changes its behavior.

What we would like to be able to do is parse the productivity effects of school choice into short-run and long-run, general equilibrium effects. Unfortunately, a neat parsing is not yet possible. Traditional forms of school choice can inform us about long-run, general equilibrium effects, but traditional forms of choice generate much weaker incentives than do reforms like vouchers or well designed charter schools. Thus, we cannot take estimates of the effects of traditional choice, subtract estimated short-term effects from a recent choice reform, and hope to use the difference as an estimate of the contribution from long-run, general equilibrium effects. There is nevertheless much we can learn from the traditional forms of choice (such as the long-term relationship between allocation and productivity effects). Thus, it is more for the sake of space than anything else that I focus on empirical evidence from recent school choice reforms, in this study.

2.3. The Endogenous Availability of Choice Options

The key obstacle for analysts is that choice options do not arise randomly, but are frequently a response to school conduct. In particular, when people are dissatisfied with a particular school’s conduct, they try to create alternative schools for themselves.

It is easy to see this phenomenon with respect to the creation of private schools, charter schools, and voucher programs. In an area where the public schools are bad, parents are frustrated and are willing to make extra effort or pay money to obtain alternative schooling. The result is an area in which private schooling is available and parents clamor for vouchers or charter schools because the public schools are bad. Recent voucher and charter initiatives illustrate this phenomenon. It is no accident that the District of Columbia has a privately-funded voucher program, a rapidly growing population of charter schools, and a nascent publicly-funded voucher program. Its public schools have historically had low productivity: its per pupil spending is in the 99[th] (highest) percentile for the US, yet its average student scores between the 10th and 20th percentiles on the National Assessment of Educational Progress. Reports of malfeasance in the District of Columbia public schools, including embezzlement, theft of school supplies, and payrolls padded with non-workers, are common.

Endogenous school choice in areas with bad public schools generates bias if a researcher naively estimates the effect of choice on productivity. Because schools with poor
productivity induce the creation of choice, it can appear as though choice causes low productivity, instead of the other way around. There are two methods that researchers can use to avoid bias: establishing pre-treatment trends and finding a rule about arbitrary assignment to treatment that mimics randomization (in the language of program evaluation, an assignment rule that is “ignorable”).

Establishing pre-treatment trends is important because it helps us predict what schools would have done in the absence of the reform. That is, it helps us with the before-after difference. In addition, pre-treatment trends help us identify a control group and use the control group’s “after” trend to predict the changes we would have seen at the treated schools in the absence of reform. If we have multiple years of data on both the control and treatment groups before the reform, we can establish their ex ante similarity and the time trends they were on ex ante.

Finding a rule about arbitrary assignment to treatment that mimics randomization is also very important. Researchers need to find control schools that were excluded from the reform for some reason that is uncorrelated with factors that affect their future performance. Such arbitrary exclusion can sometimes be found in policy rules or natural events. However, in some school choice reforms, no arbitrary exclusions exist. For instance, when Chile introduced school choice, the same law applied across the entire country. Therefore, the variation in school choice that arose was entirely endogenous. No researcher has yet identified any variation in Chile’s school choice that was arbitrarily assigned.

Studies that use both arbitrary treatment rules and pre-treatment trends are far preferable to studies that use only one of the two methods. Studies that use neither method are not credible because they analyze variation in school choice that may all arise endogenously. They are likely to claim that they have discovered the effects of school choice when they have mostly uncovered the effects of the circumstances (such as bad public schools) that caused school choice. Let us briefly return to the Chilean example because it is a situation where no arbitrary assignment exists and differential growth in voucher schools was purely endogenous. Apparently, for Chile, no pre-treatment data exists. Thus, researchers have neither pre-treatment trends nor arbitrary assignment to treatment, and none of the several studies on Chilean vouchers is sufficiently credible to be given much weight.

2.4. Identifying Reforms That Produce Competition Among Schools

As mentioned above, one cannot test the hypothesis that competition among schools will raise productivity by looking at choice reforms that fail to introduce competitive incentives. One must focus on reforms where: (a) at least a substantial share of a student’s funding follows him from his regular public school to his choice school, (b) choice schools can expand and regular public schools can shrink, (c) choice schools do not depend (financially or for operating authority) on the regular public schools with which they are supposed to compete. In addition, it is practical to focus on reforms (d) that have been in place for several years, (e) in which the regular public schools could
potentially lose more than a few percent of their students, and (f) for which \textit{ex ante} data are available.

I have identified three American choice reforms that met these criteria: vouchers in Milwaukee, charter schools in Michigan, and charter schools in Arizona. Most choice reforms failed to meet one or more criteria because they were too small, had very low constraints on enrollment (for instance, no more than one percent of local students could attend choice schools), or generated perverse financial incentives (for instance, the local district loses no money when it loses a student to a choice school). In what follows, I attempt to show the best available evidence on these three reforms.

\textbf{2.5. The Effect of Vouchers on Productivity in the Milwaukee Public Schools}

Vouchers for poor students in Milwaukee were enacted in 1990 and were first used in the 1990-91 school year. Currently, a family is eligible for a voucher if its income is at or below 175 percent of the federal poverty level, which is USD 18,400 for a family of four. In the 2002-03 school year, the voucher amount was USD 5,783 per student or the private school’s cost per student, whichever was less. For every student who leaves with a voucher, the Milwaukee Public Schools lose state aid equal to 45 percent of the voucher amount (up to USD 2,602 per voucher student in 2002-03). Milwaukee’s per pupil spending in 2002-03 was USD 11,436 per pupil, so the district was losing 23 percent of the per pupil revenue associated with a voucher student. The vouchers may be used at secular and non-secular private schools.

The voucher program had a difficult start. While approximately 67,000 students were initially eligible for vouchers, participation was initially limited to only 1 percent of Milwaukee enrollment. In 1993, the limit was raised to 1.5 percent. These limits were binding: the vouchers were oversubscribed and a lottery was held among applicants. In short, after the first small shock, the Milwaukee public schools stood to lose no further students. The voucher was worth only USD 2,500 (38 percent of local per pupil spending), and—moreover—the Milwaukee Public Schools were “held harmless” financially: the state ensured that they lost no money when a student took a voucher. Legislative and court battles kept the future of the program continually in doubt: it was typical for parents not to know, in the middle of the summer, whether their voucher would actually be usable in the fall. This situation continued up until 1998, and evidence of all sorts (statistical, from the Milwaukee superintendent, from school board members, and from political analysts) suggests that Milwaukee did not feel normal competitive pressures over this time. For a narrative, see Hess (2002), who argues that the main response of the Milwaukee Public Schools in these early years was to dedicate itself to the political campaign to end the voucher program. In the spring of 1998, the situation changed substantially after the voucher program was affirmed by the Wisconsin Supreme Court, a decision that had been very much in doubt. At that time, the voucher was raised to about USD 5,000, about 50 percent of the funding for it began to come from the Milwaukee Public Schools’ budget, and the ceiling on enrollment was lifted to 15 percent of Milwaukee students. Because the program was already established and familiar to Milwaukee residents by spring 1998, take-up of the program immediately sextupled. Nevertheless, the new enrollment ceiling did not immediately bind and any eligible
student who wanted a voucher could have one. Overall, the voucher program generated substantial competition starting in the 1998-99 school year, but generated almost no competition before that.

Not all schools in Milwaukee experienced the same increase in competition as the result of the voucher program. The greater was a school’s share of poor children, the greater was the potential competition because the greater was the potential loss of students (after 1998). Some Milwaukee schools had as few as 25 percent of their schools eligible for vouchers, while other Milwaukee schools had as many as 96 percent eligible. Also, because private elementary schools cost significantly less than private high schools, more than 90 percent of vouchers have been used by students in grades one through seven (1998 through 2003). Thus, only elementary schools in Milwaukee faced significant potential competition.

These facts about the voucher program suggest that the following type of evaluation is most appropriate for examining the productivity response of Milwaukee Public Schools. First, one should focus on the productivity of Milwaukee schools in grades one through seven. Second, schools’ productivity should be compared before and after significant competition. The clear “before” period ends in 1996-97; the clear “after” period begins in 1998-99; the 1997-98 school year straddles “before” and “after.” Third, schools in Milwaukee can be separated into those that were “more treated” by competition because a large number of students were eligible and those that were “less treated.” More treated schools are likely to have responded more strongly to the program. We can think of the less treated schools in Milwaukee as a partial control group, but all schools in Milwaukee were eligible for non-negligible treatment. Therefore, it is desirable to have a control group of schools from Wisconsin that were truly unaffected by the voucher program. There are a small number of good control schools in Wisconsin that are comparable in poverty and minority representation to Milwaukee’s schools. I chose the most similar schools available for the evaluation, but it is likely that the results will understate the productivity effects of school competition. We expect understatement because the control schools had slightly fewer poor and minority students and consequently enjoyed greater productivity and higher productivity growth than the most affected Milwaukee schools (this is a common finding and is demonstrated by the ex ante trends). That is, the control schools and the less treated schools in Milwaukee would have enjoyed higher productivity growth than the most treated schools in the absence of the reform, so using them to control for productivity growth causes me to understate the positive effect of vouchers on the more treated schools’ productivity.

[Milwaukee elementary schools can be divided into three groups]: most treated (Milwaukee schools where at least two-thirds of students were eligible for vouchers), somewhat treated (Milwaukee schools where less than two-thirds of students were eligible for vouchers[]), and untreated comparison schools. There are 32 most treated and 66 somewhat treated elementary schools. All of the Milwaukee elementary schools have enrollment of about 71-72 students in a grade. In the most treated schools, an average of 81.3 percent of students were eligible for free or reduced-price lunches (and thus eligible for vouchers), 65.4 percent of students were black, and 2.9 percent of students were Hispanic. In the somewhat treated schools, an average of 44.5 percent of students were
eligible for vouchers, 49.1 percent of students were black, and 13.7 percent of students were Hispanic. I included a Wisconsin elementary in the untreated comparison group if it (1) was not in Milwaukee, (2) was urban, (3) had at least 25 percent of its students eligible for free or reduced-price lunch, and (4) had black students compose at least 15 percent of its students. There were 12 schools in Wisconsin that met these criteria. In the untreated comparison schools, average enrollment in a grade was 51 students, 30.4 percent of students were eligible for free or reduced-price lunch (and, thus, would have been eligible for vouchers had they lived in Milwaukee), 30.3 percent of the students were black, and 3.0 percent of students were Hispanic.

Students in Wisconsin take state-wide examinations in grades three, four, eight, and ten. Because I am necessarily focusing on the productivity reactions of elementary schools, I use achievement measures based on the third and fourth grade tests. The third grade test is a criterion-referenced reading exam for which the state reports the share of students who attain various levels of proficiency. It is difficult to use proficiency shares in a measure of productivity, so I construct productivity by dividing a school’s fourth grade scores, which are expressed in national percentile points, by its per pupil spending in thousands of real (1999) dollars.

Before examining productivity or regression results, let us look at simple time-series of achievement data from the most treated, somewhat treated, and untreated comparison schools.

In the years before 1998, the most treated schools had a declining share of students performing at a proficient or advanced level and an increasing share of students were performing at a minimal level. In the 1998-99 and 1999-00 school years, these trends reverse themselves. [In] somewhat treated schools, [the time trends are similar]. In particular, the trends prior to 1997 are bad, but, in 1998-99 and after, students first become more likely to perform at the basic, instead of the minimal level, and then become more likely to perform at the proficient, instead of the basic or minimal levels. Finally, . . . there were no interesting time trends in students’ performance at the untreated comparison schools, before or after the reform.

We can confirm these third grade results with the fourth grade results, which are somewhat easier to display because they are measured in national percentile rank points. The first year of the fourth grade tests was 1996-97, and the most recent results are from 2001-02.

[T]he untreated comparison schools are doing pretty much the same, before and after the reform. In contrast, both the most treated and somewhat treated schools display dramatic improvement in 1998-99 and 1999-00, followed by a plateau. The most treated schools gain about 8 national percentile points overall. [There is an] even more sizable improvement in science achievement. Once again, the untreated comparison schools have flat achievement, while the most treated and somewhat treated schools show exceptional improvement in 1998-99 and 1999-00, followed by steady scores. The most treated schools gain about 13 national percentile points overall. Finally, . . . results for the language exam, which tests vocabulary and grammar, not a foreign language[. are]
noisier . . . . For instance, all the schools show declining scores between 1999-00 and 2000-01, probably as a result of the inclusion of test items on which poor, urban children were likely to do badly. In any case, the overall pattern is still evident: Relative to the untreated comparison schools, the most treated and somewhat treated schools perform better after the 1998 voucher reform.

. . . Regression results . . . solidify the[se] patterns . . . . [A]fter the voucher reform, the share of students whose performance [on the third-grade reading exam] was minimal declined by 9.3 and 8.4 percent at the most and somewhat treated schools, respectively. Conversely, after the voucher reform, the share of students whose performance was proficient increased by 4.1 percent at both the most and somewhat treated schools. All of the results just quoted are statistically significantly different from zero at the 95 percent level of confidence. Moreover, it is credible that these results are causal because they are measured relative to the untreated comparison schools (the omitted group), relative to the schools’ own previous level of performance, and relative to the schools’ own previous trend in performance. They are not the result of year-to-year variations in the test because there are year effects in the regression . . . .

. . . [A]fter the voucher reform, [fourth-grade] students at most treated schools scored 8.1, 13.8, and 8.0 national percentile rank points better in math, science, and language, respectively. Students at somewhat treated schools scored 5.7, 11.1, and 6.3 national percentile rank points better in math, science, and language, respectively. All of these results are statistically significantly different from zero at the 95 percent level of confidence, except for the math result for somewhat treated students, which is at the 90 percent level.

[Let us focus on] the estimated effects of vouchers on Milwaukee schools’ productivity, as opposed to achievement . . . to check that the[se] achievement gains . . . were not merely due to increases in the Milwaukee Public Schools per-pupil spending (which did rise slightly, in a mechanical fashion, with each voucher student’s departure). . . . Milwaukee public elementary schools became significantly more productive. Productivity rose by between .9 and 1.7 national percentile points per thousand dollars in the most treated schools, after the voucher reform. It rose by .7 to 1.3 national percentile points per thousand dollars in the somewhat treated schools. These increases are all statistically significantly different from zero at the 95 percent level of confidence. It is credible these results are causal because . . . they are measured relative to the untreated comparison schools, relative to the schools’ own previous level of performance, and relative to time effects.

Overall, the improvements in the Milwaukee public schools, following the 1998 voucher reform, are very impressive and have been maintained. In fact, the improvement is so impressive that people have sometimes asked whether they might not be due to severe reverse cream-skimming. That is, perhaps the voucher schools removed all the worst students from the Milwaukee Public Schools? After years of being asked about vouchers generating cream-skimming, it is refreshing (if peculiar) to be asked about the reverse. However, it is easy to show that reverse cream-skimming cannot account for the large improvements. For instance, between 1996-97 and 1999-00, voucher applicants scored
about six points lower in language and about ten points lower in math and science than the average Milwaukee student. Applicants scored at the same level as other low income Milwaukee students who were eligible for the vouchers. Thus, the voucher students’ departure would raise fourth grade scores in Milwaukee public schools by at most one point in language and two points in math and science. Also, the gains would occur for the district overall, but would not be concentrated at the most treated schools (where the remaining students were most like the departing students). Indeed, one can do an even more extreme calculation. Assume that the voucher students were the very worst students in Milwaukee prior to their departure—that is, the vouchers literally cut off the bottom tail of the Milwaukee score distribution. Even under this extreme assumption (which is far too extreme, given what we know about the scores of actual voucher takers), the departure of voucher students could not account for more than 25 percent of the actual improvement in Milwaukee public school achievement.

Overall, Milwaukee suggests that public schools can have a strong, positive productivity response to competition from vouchers. In order to get a sense of the magnitude of the response, consider the following question. Is it likely that the productivity effects of Milwaukee’s voucher program (the “rising tide”) are likely to overwhelm any allocation effects the vouchers could have? Consider the worst possible allocation change. Very high achieving Milwaukee elementary schools (top decile) score about 32 national percentile points higher in math than the lowest achieving schools (bottom decile). Thus, a Milwaukee student’s worst case scenario would be to experience a fall of about 32 national percentile points in his peer group. Moreover, let us assume that one of the highest existing estimate of peer effects is the truth. One of the highest on record in well-identified, modern studies is .3; it is from my own work and it certainly exaggerates the effect that such a peer reallocation could have. (The estimate it is not only at the top end of the estimates in the literature but is at the top end for the article in question, which generates a range of estimates from .05 to .3. Moreover, the very limited evidence that we have on nonlinearities suggests that the peer effect of a given change in scores weaken as the gap in scores grows larger. Thus, when I extrapolate linearly, I will produce an overestimate of the effect of reallocating peers.) If the student’s scores fall by 32 times .3, the improvement he enjoys from the productivity effects of vouchers will approximately cancel out the worst possible allocation effects he could experience, even assuming that he is very influenced by his peers. This is an example of the “rising tide” implications of competition, which makes it easier to contemplate allocation effects that are not easily predicted.

Section 2.6 examines the effect of charter schools on productivity in the Michigan and Arizona public schools. Hoxby concludes: Charter schools may have induced improvements in achievement and productivity in Michigan and Arizona that are positive and statistically significant, but are they large enough to matter? To put the gains in context, it may help to know that an urban district with poor students like Detroit or Phoenix would take between 10 and 20 years to reach the level of achievement that its most affluent suburbs enjoy now (it would take longer in Michigan because the achievement gap is wider). Comparing Michigan and Arizona to Milwaukee, it is tempting to conclude that productivity effects of charter school competition are in the same direction as, but weaker than, those of vouchers. Alternatively, it may be just be
that Milwaukee is a better “experiment” for evaluators because its change was more abrupt and the threat was far more targeted (remember, some Milwaukee schools could have lost nearly all their students).

3. Does Students’ Achievement Rise When They Attend Voucher or Charter Schools?

The single question most commonly asked about school choice is undoubtedly whether students’ achievement rises when they begin to attend a voucher or charter school. Despite its popularity, this question is essentially wrong-headed, for two reasons.

First, we should be asking about the productivity of schools, not achievement at schools regardless of the resources they have. If school choice is to be public policy, and not merely an experiment, then the question we need to answer is whether students’ achievement would rise if they attended voucher or charter schools that had resources like those available to them in regular public schools. In other words, we should ask the achievement question, holding resources constant (as well as holding students’ ability, motivation, and other characteristics constant). Yet, because private schools participating in voucher programs and charter schools consistently have fewer resources than public schools, researchers are forced to focus on the achievement question without holding resources constant.

The second reason why the achievement question is wrong-headed is that economic theory predicts and evidence suggests that school choice will raise the productivity of the public schools forced to compete with voucher or charter schools. Indeed, the key idea motivating school choice proponents is the expectation of positive effects on public school achievement, given the resources available. Thus, it is unclear what we are meant to do with the answer to the achievement question. Suppose that we found that school choice raised the achievement of all students, including those in the regular public schools, so that students’ achievement did not rise significantly when they attended voucher or charter schools. Surely, we would judge such a program to be more successful than one that raised the achievement of only the students who attended choice schools. (In practice, because most school choice programs have been far too small to provide meaningful competition for their local public schools, this issue has not yet posed a serious problem for evaluators. They need only avoid the small number of programs that might realistically affect achievement in the public schools.)

Finally, before looking at the evidence on the achievement question, it must be said that economists always find the question peculiar. This is because a parent is revealing his belief that a choice school is better when he continues to send his child there, rather than the regular public school his child could freely and easily attend. Suppose we were to find that students’ achievement was no better in choice schools. What would we then conclude, knowing that the parent still prefers the choice school? We might conclude that the parent valued some aspect of the school other than achievement (such as discipline or safety); we might conclude that the student’s achievement was higher on some dimension not measured by standardized tests. Given that parents observe much more than an econometrician does about his child’s schools, it would be foolish to conclude that the
parent was simply wrongheaded. In short, a parent’s continuing to choose a voucher or charter school is such a strong indication of his observations that we should hesitate to conclude that a choice school’s achievement is inferior so long as there is substantial demand for that school. (Evidence from surveys, house prices, and school choice itself suggests that parents rank schools on the basis of academics, discipline, a supportive atmosphere, and safety. I am not aware of any evidence that supports the fear that parents rank schools on some very superficial basis.)

Despite these reasons why the question—does students’ achievement rise when they attend voucher or charter schools?—is not the best question, let us consider its answer. The main challenge for researchers is that students who apply to voucher or charter programs may differ from those who do not. If students merely differed in their observable characteristics, researchers could control for their differences. But, we suspect that students who apply to choice may be different in ways that are unobservable: their parents may be more motivated, they may be getting into a bad pattern in their current school, and so on.

3.1. Evidence from the Best Available Empirical Method: Randomized Control Groups of Students

Because of this challenge, the most credible research is that in which choice students are compared to students who applied to the same choice program but who were randomly not assigned to a voucher or charter school. Such random assignment occurs when oversubscribed choice schools hold lotteries among applicants to determine who enrolls. Such lotteries are common because many choice programs are oversubscribed and randomization is often the mandated method of dealing with excess demand.

Studies of choice students and their “lottered-out” counterparts usually present a few estimators: (1) a straightforward comparison (the coefficient from a regression of achievement on a dummy for being lotteried-in); (2) the coefficient from a regression of achievement on a dummy for being lotteried-in plus observable student characteristics such as race, gender, and poverty status; and (3) the coefficient from an instrumental variables regression in which the instrument for attending a choice school is a dummy for the student’s having been lotteried-in. The first two measures, which are typically very similar, estimate the “effect of the intention to treat.” The third measure estimates the “effect of treatment on the treated.” The effect of treatment on the treated is the measure in which policy makers are most interested. It reveals the effect of attending a choice school, free from any bias due to voucher winners’ deciding whether to attend a choice school with their voucher. The effect of the intention to treat is the effect of being offered a voucher, regardless of whether the student uses it. The intention to treat estimator is useful if one thinks of the choice program as including all those to whom it was offered, not only those who use it. If the voucher winners who do not ultimately enroll in choice schools are poorer or less motivated than other voucher winners, we would expect the intention to treat estimate to be lower than the effect of treatment on the treated. On the other hand, if the voucher winners who ultimately do not enroll in choice schools are those whose prospects are best (parents tend not to switch schools when their child is
succeeding), we would expect the intention to treat estimate to be higher than the effect of treatment on the treated.

... [In studies, t]he effect of the voucher on black students who actually use it (effect of the treatment on the treated) ranges from a low of 4.3 national percentile rank points after two years in New York to a high of 9.0 national percentile rank points after two years in Washington, the District of Columbia. Roughly speaking, the differences among the cities are consistent with the quality of their high poverty schools. In other words, one might tentatively conclude that private (voucher) schools produce quite similar achievement in poor students everywhere but that the quality of high poverty public schools varies with the district. For instance, Washington’s high poverty public schools are worse than those of New York (based on simple comparisons of test scores), so vouchers produce greater achievement gains in Washington. In [some studies], the effect of treatment on the treated is slightly greater than the effect of the intention to treat. In [a Milwaukee study], this difference is reversed. All this suggests that the bias due to students electing not use their voucher varies with the program design and environment.

[In] the publicly-funded voucher programs in Cleveland and Milwaukee[,] the effect of treatment on the treated is 11 to 12 national percentile rank points in mathematics and 6 national percentile rank points in reading (after only two years in Cleveland and after four years in Milwaukee). These estimates are for all students in the sample, about 80 percent of whom are black.

An apparently striking result is that the gains appear to be restricted to black students or groups largely composed of black students. This result may be an artifact of the location of the voucher programs that have been evaluated. All of them are in cities in which the poor student population is predominantly black. Moreover, the eligible black students are concentrated in certain neighborhoods. If they are lotteried-out, they attend a common set of public schools. If they are lotteried-in, they find a common set of private schools nearby. In contrast, these cities’ eligible non-black students are more idiosyncratic in location, neighborhood quality, default public school, and proximate private schools. For instance, if we use block-group level 2000 Census data . . . , we find the following geographic concentration indices: for poor black school-aged people, .0112; for poor white school-aged people, .0008; for poor Hispanic school-aged people, .0016; for poor Asian school-aged people, .0019. That is, the poor black school-aged population is 14 times more concentrated geographically than the corresponding white population, 7 times more concentrated than the corresponding Hispanic population, and 6 times more concentrated than the corresponding Asian population. In short, the non-black student population is not only generally small in the voucher samples, it is much less homogeneous in its school experience than the black student population. The combination of smaller sample size and greater noise is likely to produce results that are not statistically significantly different from zero, even if there is a true effect. In short, we would be unwise to conclude that vouchers have zero effect on non-blacks; a more reasonable interpretation of the evidence is that researchers will not discover how vouchers affect them until there is a voucher that targets a concentration of them. (Put another way, there is a difference between a precisely-estimated zero effect and a effect that is statistically insignificantly different from zero.)
Results similar to those shown in the New York City [study] have been criticized by Krueger and Zhu (2003): Howell and Peterson (2003) have responded to the criticisms at length. Krueger and Zhu argue that the New York City results are “so sensitive” to changes in the definition of black ethnicity “that the provision of vouchers in New York City probably had no more than a trivial effect on the average test performance of participating Black students.” Krueger and Zhu made two main changes to the data: they included students without baseline scores (primarily kindergarteners) and they recoded students’ racial and ethnic identities. The first change, in and of itself, did not significantly alter the results, so it is not worth discussing further. It is only by combining the first change with the second change that Krueger and Zhu generate statistically insignificant results with the New York City data.

Howell and Peterson use the United States Census definition of Hispanic and use three standard methods of identifying a student’s race: self-identification, race of the custodial parent, and race of either parent. With all of these classifications, they obtain results similar to those shown in the New York City [study]. Krueger and Zhu, in contrast, reclassify as black some people whom the Census defines as Hispanic. They also classify mixed-race students who have a black father as black, even though they use only the custodial parent’s race (usually the mother’s race) for all other mixed-race students. It is standard social science practice for a researcher to maintain an arms-length relationship between his creation of variables and his results. That is why we adopt Census or other standard definitions of many variables: it makes it impossible for us to adjust the construction of a variable to generate a particular result. That is also why we prefer to treat groups symmetrically when constructing variables. Once we allow ourselves to create novel classification schemes for standard data and to use a different method of classification for each group, we have so many degrees of freedom that our decisions control our results. In short, social scientists have developed protocols to unmask specification searching. The Krueger-Zhu classification scheme has never been used before, arbitrarily violates standard classification of Hispanics, and deals asymmetrically with blacks. With the wide latitude for specification searching they allow themselves, it is not surprising that Krueger and Zhu find a particular specification that generates statistically insignificant results. This could not be described as a situation, however, where results were highly sensitive to reasonable variations in the specification.

3.2. Achievement Gains May Be Concentrated Among Students Particularly Likely to Be Exploited Now

I have suggested that there may be an econometric interpretation of the difference in results for black students, but a structural interpretation may also be correct. We may speculate that, owing to discrimination in the housing market, poor blacks are less able to exercise Tiebout choice (choosing a school by choosing a neighborhood) than non-blacks. We may also speculate that, owing to discrimination by teachers and administrators, staff feel more comfortable underserving students and taking rents when they work in schools that serve black students, as opposed to non-black students. Under either scenario, vouchers would plausibly constitute a greater positive shock to blacks’ choice sets than they would to non-blacks’ choice set.
3.4. Remembering That It Is Productivity That Matters

At the outset of this section, I stated that it was wrongheaded to make simple achievement comparisons when productivity comparisons are what we need for policy making. Thus, before leaving this section, let us consider what we have learned about the relative productivity of public schools and choice schools. Indeed, for the sake of this comparison, let us focus on the result that the achievement of students at choice schools is certainly no lower on average than it would be at public schools. If we say that achievement is roughly the same, then the difference is productivity is a function of the difference in average inputs. For instance, in the five cities [whose studies were reported above], the public schools spent an average of USD 9,662 per student and the voucher schools spent an average of USD 2,427 per student (this is spending, which is greater than tuition). These spending numbers, combined with achievement that we will call equal, suggest that the voucher schools were 298 percent more productive. When interpreting this number, remember that we have already controlled for differences in student ability and motivation through the randomization. Also remember that voucher students were never the richer and easier-to-educate students in the public schools. Even if we think that the 298 percent measured difference in productivity is somewhat off, it is very unlikely that the true productivity difference is zero or small.

4. Do Voucher and Charter Schools “Cream-Skim”?

Economic models predict that school choice programs can affect the allocation of students among schools in a wide variety of ways, depending on assumptions about (a) the amount of money that follows a student, (b) the relationship between a student’s voucher (charter school fee) and his characteristics (family income, special education status, and so on), (c) whether the voucher (charter school fee) is affected by the characteristics of the student’s neighborhood, sending school, or receiving school, (d) whether choice schools can exercise selection, (e) the political process that determines how the loss of students to choice schools plays out in government support for regular public schools, (f) the relationship between the local housing market and support for regular public schools (which usually depend on property taxes), and (g) the importance and functional form of peer effects. There are no general predictions about allocation effects. Almost any interesting and plausible allocation equilibrium can be generated with the right set of assumptions about factors (a) through (g) above.

4.1. Approaches to Dealing with the Multiplicity of Allocation Outcomes

There are three useful approaches that economists can take for dealing with this great multiplicity of possible allocation outcomes. The first is to admit that theory usually offers us a array of predicted outcomes and to therefore proceed in a purely empirical fashion, describing the allocation outcomes we see and attempting to identify patterns when possible. The second is to design choice programs so that they satisfy certain assumptions and produce a desirable set of allocation outcomes. The third is to derive outcomes for a set of assumptions that are as realistic as possible for parameters and relationships where we have evidence (for instance, making assumptions about the initial distribution of students or house prices based on actual data) and to test a wide variety of
assumptions for parameters and relationships on which we have little or no evidence (for instance, the importance and functional form of peer effects).

Some economists have taken a fourth approach, which is to derive outcomes for a set of assumptions that are obviously unrealistic but that generate tidy or dramatic outcomes. The kindest way to look at such approaches is that they were a necessary, early stage of investigation that helped us to develop our tools and appreciate the multiplicity of allocation outcomes. For instance, in a paper early in their agenda, Epple and Romano (1998) derived outcomes for a model in which there was only one public school (so that the public sector mechanically had no sorting), there were a great many private schools (allowing tremendous opportunity for sorting), and in which peer effects on achievement always satisfied single crossing (a higher ability student always benefits more from getting another high ability peer than a lower ability student does, no matter what the initial distribution of peers). Moreover, in their model, the public school lost its entire per-pupil spending when a student took a voucher despite the fact that the voucher was only a small share of per-pupil public spending! Each one of these assumptions that can be compared with reality is grossly unrealistic. The assumption that cannot be compared with reality is the one about peer effects, and it is a very restrictive assumption. We have no evidence on which to base such strong functional form assumptions. Such papers undoubtedly helped economists develop their tools for analyzing school choice, so we must be grateful that they were written. However, in retrospect, it is difficult to do much with their results. After all, there are an infinity of sets of unrealistic assumptions. Why should we privilege one set of unrealistic assumptions over the infinity of others?

In this paper, I will take a variant of the first, or purely empirical, approach. Cream-skimming is not a general prediction of choice models and tends to occur in models in which the public school system allows very little sorting, voucher eligibility is broad, vouchers are uniform in size, and peer effects exhibit single-crossing. Nevertheless, the common person is more likely to worry about cream-skimming than about other allocation outcomes. So, I will privilege this potential outcome and look for patterns of it in the data from actual reforms.

4.2. Evidence About Cream-Skimming from Patterns of Race, Ethnicity, and Poverty

In practice, it is quite easy to look for these patterns on dimensions like students’ race and poverty (see below for achievement). This is because we can observe the race and poverty of students who participate in choice programs, observe the race of poverty of students who remain in the schools from which the programs draw, and reasonably assume that students’ race and poverty status do not change over time. Such investigations are imperfect—factors other than choice reforms may affect the racial and poverty composition of a regular public school and children in choice schools would not necessarily attend the regular public schools in the absence of the reform. Nevertheless, we can learn something from these simple comparisons.

... [F]or all of the charter schools in operation in the 2000-01 school year[, one can calculate] the odds ratio that a student is black, Hispanic, or poor, relative to the district in which his charter school is located... or relative to the public school that is physically
nearest to the charter school . . . . The odds ratio is the ratio of the probability that the student is, say, black, given that he is enrolled in the charter school, to the probability that he is black given that he is in the district (nearest public school). An odds ratio equal to one means that the charter school draws its students in proportion to their proportions in the underlying population. An odds ratio greater than one means that the charter school disproportionately draws, say, black students.

. . . [T]he odds ratio is smaller than one for white students and Asian students; substantially larger than one for black students; slightly larger than one for Hispanic students; and substantially larger than one for poor students. For instance, consider the odds ratios for black and poor students. The numbers in the district column mean that a charter school student is 2.28 times as likely to be black and 1.12 times as likely to be poor than a randomly drawn student from his district. The number in the nearest school column means that a charter school student is 1.38 times as likely to be black and 1.09 times as likely to be poor than a randomly drawn student from the nearest public school. In short, . . . charter schools [do not seem to be] cream-skimming in any conventional racial, ethnic, or economic way. They are disproportionately drawing students who have suffered from discrimination, not enjoyed undue preference, in the public schools.

4.3. Evidence About Cream-Skimming from Longitudinal Data on Achievement

We would like to look for evidence of cream-skimming on the ability dimension. For data reasons, this is possible only for some reforms. The difficulty is that students’ achievement obviously can change, so we need longitudinal achievement data from both the choice school and the regular public schools, and we need to be able to link these data so that we can see what voucher or charter schools applicants were like relative to their public school peers before applying. A few states do maintain such information. Here, I rely on evidence from Chicago-area and Texas charter schools.

. . . Hanushek, Rivkin and Kain (2002) . . . look at future charter school student’s performance while in the regular public schools. They control for school times grade indicator variables, so that a student is being compared to others in his grade in his public school. They use performance on the Texas Assessment of Academic Skills, standardized so that the each grade’s mean score is zero and its standard deviation is one. They find that future charter school students do .14 standard deviations worse than their peers in reading and .30 standard deviations worse than their peers in math.

Using a similar strategy, I examined future charter school students in Chicago. I examine their annual gains on the Iowa Test of Basic Skills relative to their peers in the same school and grade. Note that I am showing their pre-charter school trajectory, not merely their level, relative to their peers. . . . I find that, prior to charter school application, future charter school students’ annual gain is 20 percent smaller in mathematics and 30 percent smaller in reading than their peers’ average gain.

Overall, it appears that choice schools are not cream-skimming in the US. The evidence mentioned above on Edison Schools confirm this. If anything, choice schools are
disproportionately drawing students who are generally considered to be less desirable or who are already experiencing achievement problems.

A reader might reasonably respond to this evidence with: “Of course the choice schools are drawing students who are minorities, poor, and low achieving. This is because the programs are designed to make them eligible and because students who are doing well in their regular public school are not going to switch schools when they typically take only part of their funding with them.” I agree and emphasize that this was my original point. We do control the allocation effects of choice programs when we design them. A multiplicity of outcomes are available. Only by ignoring both theory and evidence could we believe that a single allocation outcome, such as cream-skimming, is a general outcome of school choice.

5. Final Thoughts on School Choice and Competition

Using data from American school choice programs, I have attempted to answer a few basic questions on school choice. A wealth of important and complex questions remain. Some of these questions may be answerable with data from school choice programs from around the world. The more variation we see in program design, the better we can investigate complex questions about finance, sorting, peer effects, and the supply of choice schools. No program can contain the full variety of features, so we need to be able to learn about one feature here and another there. Nevertheless, if we are to make progress, researchers must subject themselves to the discipline of clearly describing the structure of programs, the incentives they generate, and the environment in which they operate. Only if we describe programs with measures that are clear and relatively universal can we aggregate up evidence from many studies. The self-discipline of researchers will largely determine whether the analyses of the next several years leave us with a muddle of evidence or greatly increase our understanding of school choice.

One decade’s experience of school choice has, however, allowed us to learn a good deal. Evidence from these first-generation school choice programs has answered simple questions like whether students’ achievement improves when they attend choice schools (apparently, yes, for the typical student eligible for choice programs now), whether public schools can respond to competition constructively (apparently, yes), and whether choice schools do cream-skimming (no, for programs designed as existing choice programs are). These answers should give us the confidence to design second-generation programs that are larger, better financed, and more ambitious in tackling issues like compensatory funding and varying vouchers with student and school characteristics.


Among the many reforms proposed for K-12 education are changes in governance that would increase the power of parents to choose schools and thereby make the education system function more like a market. Within this set of reforms, which also includes

Draft — do not distribute!
offering greater choice among public schools and the opportunity to establish public “charter” schools, school voucher programs are particularly controversial because they would permit parents to use public funds to secure education not only at public schools, but also at private schools. Proponents and opponents disagree about the effects of voucher programs on student achievement, on the social and racial segregation of students and on disadvantaged students. In addition, they differ on the importance of maintaining the separation between religious private schools and the state.

School voucher programs currently exist only on a small scale in the United States. The main publicly funded voucher programs are in Milwaukee, Cleveland and Florida. In addition, small privately funded programs provide vouchers for low- and moderate-income students in cities such as New York City, Dayton, Ohio, and Washington, D.C. Another privately funded program, the Children’s Scholarship Fund, operates at the national level.

Recent studies based primarily on U.S. evidence typically conclude that the data are insufficient to draw clear conclusions about the net effects of vouchers on academic achievement, access to schools, racial integration and civic education (for example, Ladd and Hansen, 1999; Levin, 2001; Campbell and Peterson, 2001; Gill et al., 2001). In the light of the limited U.S. experience, some authors support investments in large-scale voucher experiments as a way to generate more definitive information on their effects. However, before making such investments, it would behoove U.S. researchers and policymakers to pay more attention to the evidence from large-scale programs in other countries such as Chile and New Zealand. Chile, for example, has had a universal school voucher program since the early 1980s that has been subject to careful evaluation (Hsieh and Urquiola, 2002; McEwan and Carnoy, 2000; McEwan, 2000a; Gauri, 1998). In addition, New Zealand introduced in the early 1990s what some observers have referred to as a universal quasi-voucher system. Parents can choose any school within the public sector, which has included religious schools since the 1970s, and school funding is based largely on student enrollment (Fiske and Ladd, 2000).

In this paper, I marshal the available evidence, including the international evidence, to show that contrary to the claims of its proponents, a large-scale universal voucher program would not generate substantial gains in overall student achievement and that it could well be detrimental to many disadvantaged students. The case for a more narrowly targeted means-tested voucher program is stronger, but even with careful attention to its design, such a program should at most serve as one element of a broader strategy designed to provide more options and better education for disadvantaged students.

**RELEVANT CHARACTERISTICS OF THE U.S. EDUCATION SYSTEM**

Four characteristics of the U.S. education system are especially relevant to the voucher debate. First, the existing system biases parental choices toward the public sector. Second, many middle- and upper-income families currently have much more choice among schools than do low-income families. Third, K-12 education is compulsory. Fourth, parents judge the quality of schools in part by the characteristics of the students in the school.
An Educational System with a Strong Bias Toward Public Production

Because all families have access to free tax financed schools at the K-12 level, they face strong financial incentives to choose public over private schools. Nationally, 12 percent of students are enrolled in private schools, the bulk of which have a religious affiliation. A generous universal voucher program could potentially eliminate this bias toward the public schools.

In many other countries, privately owned and operated schools receive large amounts of direct financial assistance from the state (Plank and Sykes, forthcoming). In contrast, this country’s commitment to the separation of church and state, as embodied in the establishment clause of the U.S. Constitution, has historically kept public funding from being used to support private schools. That situation could change now that the U.S. Supreme Court has decided that there is no constitutional barrier to the use of indirect funding for religious schools through school vouchers in Zelman v. Simmons-Harris, a July 2002 decision relating to Cleveland’s voucher program. However, widespread extension of voucher programs to religious schools must still overcome the hurdles of establishment clauses in many state constitutions and strong political opposition in many state legislatures. If one believes that the separation of church and state in connection with elementary and secondary education has provided substantial gains to the United States in the form of a more pluralistic and open society, then weakening this separation through the extension of voucher funding to religious schools could potentially generate large social costs. However, this tradeoff is one to which economists—or at least this economist—do not have much to contribute.

There’s another tradeoff as well, this one for the private schools themselves. The international experience indicates that widespread public funding of private schools typically brings with it greater regulation (Plank and Sykes, forthcoming). Such regulation is a natural way for the government to assure that taxpayers’ dollars are being used to promote the ends that justify public funding. To the extent that a universal voucher program in the U.S. context were accompanied by additional state control over areas such as curriculum, assessment and school admissions policies, many existing private schools could choose not to participate.

Choice Among Schools Is Available, but Limited

Within each school district, a child typically attends the school assigned to children in that neighborhood. Hence, a family’s decision of where to live plays a large role in where a child attends school. To the extent that middle- and high- income families congregate in well-to-do suburbs, their children tend to be grouped in suburban schools, financed in part by taxes from large local property tax bases. Conversely, students whose families are restricted by their low income or by their race to economically or racially isolated areas of central cities will also end up grouped in particular schools, often those with high concentrations of disadvantaged students, insufficient resources, low average achievement and high dropout rates.
Families with sufficient income can move among suburban school districts to increase the quality of the education their children receive (Hanushek, Kain and Rivkin, 2001) or else consider private schools. Moreover, the capitalization of differential school quality into housing prices, so that homes in school districts with higher-quality schools cost more, exacerbates the effects of income differentials (Black, 1999; Figlio and Lucas, 2000). While children from low-income families in big cities do switch schools quite frequently within districts, those moves are largely driven by the vagaries of the low-income housing market and typically do not result in higher-quality schools and better educational outcomes (Hanushek, Kain and Rivkin, 2001). Hence, a carefully designed school voucher program targeted to low-income families could potentially provide low-income families with some of the freedom of school choice now available only to higher-income families.

Recent theoretical work has highlighted another potential benefit of a school voucher program, namely, the reduction in residential segregation by income (Nechyba, 1999, 2001). The notion is that if families had greater access to private schools, they would be less likely to separate their residences by income. Based on a theoretical model calibrated to New Jersey data (a state with many small school districts and over 20 percent of its children in private schools), Nechyba’s (2001) simulations indicate that access to private schools reduces the variation in incomes across his three stylized districts by more than half and substantially increases the variation in income within districts, particularly within the poorest district. Unfortunately, the model does not incorporate racial preferences that could influence both school and residential choices. Moreover, school voucher programs would be a rather indirect policy tool for reducing residential segregation. Other options would include zoning and housing policies, as well as the expansion of parental choices among public schools so as to break the link between where families live and where their children go to school. In addition, any benefits of lower residential segregation achieved through a school voucher program would need to be weighed against less desirable outcomes that emerge from Nechyba’s stylized model. These include greater economic homogeneity within schools and lower-quality public schools in the poorest district.

Compulsory Schooling

School attendance through age 16 in the United States is compulsory. This public commitment to schooling, which reflects perceptions of large social benefits above and beyond the private benefits to the students themselves, need not require public ownership of the schools and, thus, is not inconsistent with a voucher program. Nonetheless, it does imply a much higher level of public interest and responsibility than is the case in other sectors to which K-12 education might be compared, such as higher education.

The contrast with higher education is important given the apparent success of the U.S. higher education system with its mix of public and private universities and colleges competing for students. The elite—and expensive—public and private institutions serve many students of high ability very well. At the same time, the system is highly stratified, with large differences in the academic ability of students across institutions and significant homogeneity along various dimensions, such as religion, within institutions.
Moreover, the sector falls far short of serving all college-age students. In 1998, less than 40 percent of those between the ages of 18-21 were enrolled in postsecondary education institutions, and as of 1999, only 25 percent of the population aged 25 and over had completed four years of college (National Center for Education Statistics, 2001, Tables 8 and 39). At best, the U.S. system of higher education can serve as a model for the top of the ability and wealth distribution of K-12 students.

The compulsory aspect of K-12 education also distinguishes it from the private sector of the economy. Competition works in that sector in part by the expansion or replication of successful firms and the shutting down of unsuccessful firms. In K-12 education, successful schools have few incentives to expand, especially when expansion entails admitting more costly-to-educate students. Moreover, the establishment of new schools takes time and, as has become clear from the U.S. experience with charter schools, requires upfront funding for planning and capital facilities. At the same time, in a country with compulsory education, failing schools can be shut down only if there are adequate places for their students in other schools.

**Role of Peer Groups in Family Choices**

Parents behave as if the peers of their children matter. Evidence from studies around the world indicates that parents exercising choice seek to move their children to schools in which the average socioeconomic characteristics or nonminority share of the students is higher than it would be in their original or assigned school. This phenomenon has been documented in systems as diverse as New Zealand (Fiske and Ladd, 2000; Ladd and Fiske, 2001a), Chile (McEwan and Carnoy, 2000), Scotland (Willms and Echols, 1993) and Chicago (Cullen, Jacob and Levitt, 2000).

This behavior is consistent with many parental motivations, including the quest for better schools. For example, parents might use the socioeconomic level of the parents of other children in the school as a proxy for school quality, based on the well-documented observation that the average achievement of students within a school is highly correlated with the socioeconomic and racial composition of the student body. Not only are levels of achievement higher in such schools, so are educational gains in each grade. Data from North Carolina and other states show, for example, that the schools with larger gains in test scores are those with higher proportions of white and higher proportions of nonpoor students (Ladd and Walsh, 2002; Clotfelter and Ladd, 1996).

The positive correlation between the socioeconomic composition of a school and the performance of its students largely reflects what happens at home rather than at school. However, school-related factors may also help to explain why schools serving more affluent and nonminority students tend to exhibit larger gains in test scores than those serving more disadvantaged students. Schools serving more affluent students may benefit from positive peer or spillover effects from one student to another; they can more easily maintain educational processes, such as assigning homework; they are more able to attract high-quality teachers; and they typically have access to more resources in the form of both budgetary resources and those provided by parents in the form of contributions and volunteer activities (Fiske and Ladd, 2000).
This observation that the “customer mix” matters to parents has four substantial implications for all educational systems, including systems financed by vouchers. First, it generates a hierarchy of schools. In the context of an abstract model in which parents judge school quality in part by the average ability of the students in the school and in which private schools charge tuition, Epple and Romano (1999) show that students with the lowest ability and lowest family income end up concentrated in public schools at the bottom of the hierarchy. Other students are distributed among a set of private schools that differ from each other by the ability and income of their students, with the private schools at the top attracting the most able and most affluent students. Empirical work based on U.S. data generally supports the model’s predictions (Epple, Figlio and Romano, 2000).

Second, when the characteristics of the school’s student body are an important determinant of the school’s quality, no simple programs or educational strategy can make a school with a large proportion of disadvantaged or low-performing students look effective. In many instances, the best strategy for such schools is to try to raise the quality of their student intakes, a strategy that cannot work in the aggregate.

Third, successful schools will be reluctant to expand if doing so requires lowering the average socioeconomic or ability level of their students. In New Zealand’s experience with full parental choice and self-governing schools, successful schools in urban areas had no desire to expand their enrollment. To the contrary, they did everything they could to maintain the mix of students that made them attractive to parents and students in the first place (Fiske and Ladd, 2000).

Finally, schools with large concentrations of disadvantaged students have difficulty competing for students (Ladd and Fiske, 2001a). This observation does not, by itself, rule out vouchers as a policy tool. For policymakers concerned about equity, however, it raises some warning flags. As discussed further below, it casts serious doubt on the proposition that competition will improve the schools serving students who attend schools at the bottom of the distribution.

**Effects of a Voucher System on Achievement and Productivity**

Improving student achievement is typically the single most important goal of current education reform efforts. A large-scale voucher program could potentially affect student achievement through three interrelated mechanisms (Hsieh and Urquiola, 2002). First, it would shift students from the public sector to the private sector. Provided the private sector were more productive than the public sector in generating student achievement, this sectoral shift would increase the productivity of the education system. Second, such a program is likely to generate greater socioeconomic and racial polarization of students among schools as students seek to improve the quality of their peers. This greater polarization may increase overall achievement, decrease it or have no effect depending on how one’s peers affect the achievement of different groups of students. Third, the introduction of a voucher system would increase competition for students. Such competition, proponents argue, would increase achievement by forcing the public schools to become more effective. The following sections look at the issues and evidence related to these mechanisms.
The relative productivity of public and private sector schools is the most well studied component of the three mechanisms. The two major strands of the research, however, have relatively little to say directly about a large-scale voucher program.

One strand of the research was initiated by James Coleman and others in the early 1980s, using national data sets such as High School and Beyond or the National Educational Longitudinal Survey. In their seminal study, Coleman, Hoffer and Kilgore (1982) conclude that students in private high schools, most of which were Roman Catholic, outperformed their public school counterparts. However, that study did not fully account for differences in who enrolls in Catholic schools, and so many subsequent researchers have reexamined the issue paying close attention to the problem of self-selection.

To that end, researchers have had to grapple with the statistical challenge of finding an appropriate instrument, that is, a variable that is correlated with a family’s decision to choose a Catholic school, but is not a direct determinant of educational outcomes. Many authors have used some function of whether a student is Catholic or the proportion of the county that is Catholic for this purpose. Such studies tend to show that Catholic schools appear to have at most small effects on student achievement as measured by test scores, but somewhat larger positive effects on the probability that students will graduate from high school and will attend college (Evans and Schwab, 1995; Neal, 1997; Grogger and Neal, 2000). In general, the benefits seem to be largest for urban minorities.

Other researchers have criticized this statistical approach on the grounds that Catholicism may be a direct determinant of educational outcomes, and they have proposed alternative measures. In their study of the relative productivity of all private schools, for example, Figlio and Stone (1999) use a set of instruments related to state labor laws, maintaining that such policies affect residents’ perceptions of the differences in performance between the public and private sectors and hence their inclination to use the private schools. Although Figlio and Stone report somewhat smaller positive overall effects on educational attainment than those that emerge from the studies of Roman Catholic schools, they do find higher test scores in math from attending religious schools for the subgroup of African-Americans in big cities.

A second strand of the research uses the recent voucher experiments to investigate the productivity of private elementary schools compared to that of public schools. Most of the knowledge about the effects of private schools at the elementary level emerges from evaluations of the publicly funded Milwaukee Parental Choice program and of the privately funded programs in Dayton, Ohio, Washington, D.C. and New York City. All of these voucher programs serve only a small fraction of the eligible students. Until the late 1990s, the Milwaukee program served no more than 1.5 percent of the district’s 90,000 students. During the time covered by the studies, the New York City program offered scholarships to only 1,300 students, the Dayton program to 515 public school students (and to another 250 already enrolled in private schools) and the Washington, D.C., program to 460 students (Peterson et al., 2000).
Results from the publicly funded Milwaukee program have been the most controversial, in large part because that program was not set up as a true randomized experiment and so researchers had to exercise judgment in choosing the appropriate control group to which the voucher students would be compared. Witte, Stern and Thorn (1995) compared the performance of voucher students to that of a random sample of other students in the Milwaukee public schools and concluded there were no significant achievement gains for voucher students. Greene, Peterson and Du (1998) compared voucher students to potential users of vouchers who were not admitted to their preferred schools and concluded that by the third and fourth years of the program, voucher students exhibited significant gains in both math and reading (of the order of 0.1 to 0.5 standard deviations). Looking at the same data a third time, Rouse (1998) was able to build on and improve upon the research methodologies of the two previous studies; for example, she took into account the fact that many students who were given vouchers did not exercise their right to use them and also paid attention to nonrandom attrition of students from the sample over time. Rouse’s sensitivity to the statistical problems provides some confidence in her finding that the program generated small gains for students in math, but none in reading.

Recent studies of privately funded voucher programs in Dayton, Ohio, Washington, D.C., and New York City provide additional information on how voucher programs—and hence private schools—affect the achievement of elementary and middle school students. In contrast to the Milwaukee program, each of these programs was set up as an experiment with random assignment of children. Participating families filled out baseline surveys of background information and, in principle, all children in both the treatment and control groups were tested annually. In contrast to Milwaukee, the private funding of these programs made it possible for students to use their vouchers in religious as well as secular schools. The programs limited eligibility to families with low income (New York City) or low and moderate income (Washington, D.C., and Dayton). Voucher amounts were generally below $2000 per child (Howell and Peterson, 2002).

Based on three years of the voucher programs in New York and Washington, D.C., and two years in Dayton, Howell and Peterson (2002) find no evidence of a general achievement difference between the public and the private schools. In no year and in no individual city (other than the second year in Washington, D.C.) was there evidence that students who shifted to private schools achieved at higher average levels than students who remained in the public school system. Further, when the analysis was disaggregated by the race of the students, no differences emerged for either white or Hispanic students. Only for the subgroup of African-American students did positive differences in achievement emerge. Even for this group, however, the differences were consistent across neither cities nor grades. For example, African-American students in Washington, D.C., who shifted to private schools achieved at far higher levels in the second year of the program, but their gains were negligible in years one and three. Although the New York study generated a more consistent average pattern of achievement over time, the positive differential emerged clearly and consistently only for students in the fifth grade (Howell and Peterson, 2002, Table 6.2, Table D.1).
Based on Howell and Peterson’s (2002) preferred estimates, which disproportionately weight the relatively stable New York results, African-Americans who switched to private schools scored about 3.9, 6.3 and 6.5 percentile points higher than did comparable students in the control group in the first three years of the program. These effects are based on the national distribution of percentile rankings on the Iowa Test of Basic Skills, and, for perspective, the gains are about two-thirds the size of the differences that emerged in another intervention that helped African-American students—the Tennessee experiment that reduced class sizes.

These effects for African-Americans represent estimates not of the offer of a voucher per se, but rather of the shift to a private school. This distinction is important given that only 53 percent of the students offered vouchers in New York City and less than 29 percent in Washington, D.C., were still in private schools three years into the program and that the students who used the voucher to attend private schools were not a random sample of those offered a voucher. In New York City, for example, having a mother on welfare reduced the probability that a student would accept a voucher by 8 percentage points and the probability of remaining in a private school for two years by another 7 points, and having a learning disability reduced the probabilities by 14 and 13 percentage points. Working in the other direction, being religiously observant significantly increased the probabilities (Howell and Peterson, 2002). Thus, the positive differential effect for African-Americans attending private schools applies not to a random sample of low-income African-Americans, but rather to a particular subset of this group.

Furthermore, it is not at all clear that this positive effect of private schools can be extrapolated to an expanded voucher program, even one targeted at a similar group of African-American students. The issue turns on the explanation for the observed achievement gains. If the apparent success of private schools for African-American students reflected the autonomy and absence of bureaucracy in such schools, then, following the logic of Chubb and Moe (1990), an expanded private school sector could, in principle, generate comparable gains. However, the observation that the differential achievement gains emerged only for African-Americans argues against this explanation. Alternatively, if the success of the private schools reflected a better match between the needs of African-American students and the offerings of particular schools, then expanding the private sector would once again continue to generate comparable positive benefits, but only if the factors that parents were looking for could be replicated in newly established private schools. The problem is that the measured success of the African-American students who shifted to private schools likely reflects in part the more disciplined student bodies in the private schools, especially in the Catholic schools, or the more advantaged and motivated group of students whose parents were willing to pay the private school tuition. These school characteristics cannot be replicated as easily in new private schools. This explanation for the positive findings in the Howell and Peterson (2002) research cannot be ruled out, since the authors had no data on the mix of students in the receiving private schools.

More explicit evidence about the sector effect for a large-scale voucher program emerges from Chile, where the universal voucher program generated a large number of new for-profit secular private schools that operated alongside the more established and somewhat
better resourced Catholic schools. Careful analysis of fourth-grade achievement data in Chile indicates that, compared to the traditional public schools, Catholic schools generated higher achievement in Spanish and math while the new secular schools produced marginally lower achievement in Santiago and even lower achievement outside the capital city (McEwan and Carnoy, 2000). This observation is important to the U.S. debate, where advocates of vouchers tend to use evidence from some types of private schools, namely Catholic high schools, or a combination of existing Catholic and other schools in the voucher experiments, to generalize to an expanded private sector that would inevitably include many types of private schools, not all of which would be able to attract the same mix of students as in the existing private schools.

Even if private schools do not generate positive differences in student achievement, lower costs could still make them more productive than public schools. Unfortunately, it is difficult to compare the true economic costs of education at private and public schools. One complicating factor is that public schools typically serve a greater proportion of students who need costly services such as special education or vocational education. Another is that tuition payments, or even total expenditures by private schools, do not represent their true costs (Levin, 1998). Private schools, especially religious private schools, receive resources in many forms: special fees, church subsidies, teachers working at below-market wages and donations of money, time, land and buildings.

If private schools could operate more cheaply than public schools, one would expect for-profit education firms like Edison or Tesseract to make a profit from taking advantage of those cost efficiencies. The fact that such firms have not been making profits suggests that cost efficiencies from private production are illusory. Further suggestive evidence comes from the voucher experiments. For example, the private schools in New York City serving the African-American voucher students offered significantly smaller class sizes than did the public schools (Howell and Peterson, 2002). Given the large share of education costs attributable to teachers and classrooms, this observation suggests that the true resource costs of educating students in private schools could well exceed those of the public schools serving comparable students.

Thus, one should expect neither higher overall achievement nor lower resource costs as a result of a shift of students from public to private schools. At most, there are likely to be small achievement gains for a selected group of African-American students. Furthermore, a universal voucher program could possibly require the government to spend more public funds on education, because some of the voucher funds would undoubtedly go to families who would otherwise have paid all of the cost of putting their children in private schools.

*The Peer Effect*

Because many parents use the social and ethnic composition of a school’s students to judge a school’s quality, a large-scale voucher program—or, more generally, any unrestricted educational choice program—is likely to increase the racial and socioeconomic stratification of schools. Other aspects of voucher programs would also contribute to stratification to the extent that they placed low-income families in a less favorable position to exercise choice than higher income families. For example, low-
income families would be disproportionately affected if the government did not pay for transportation to the chosen schools, if voucher schools were allowed to charge fees and tuition in addition to the amount of the voucher, if schools were allowed to select their students or if low-income families have less access to information than did high-income families. While a voucher program could be designed to mitigate these effects, the more basic pressure for stratification would remain as long as the voucher program were not restricted to low-income households.

The question here is how increased stratification is likely to affect the overall productivity of the education system. A variety of studies in which the authors have carefully addressed various thorny statistical problems have found evidence of pure peer effects in the sense of spillovers within the classroom from one student to another (Hoxby, 2000b; Hanushek et al., 2001). Positive peer effects also emerge from other studies, but not always consistently, as in Henderson, Miezkowski and Sauvageau (1978), Willms (1986), Zimmer and Toma (2000) and Bryk and Driscoll (1988).

If peer effects were positive and linear, the gains in achievement for the students who move out of the public schools in search of higher-quality peers would be exactly offset by the losses to other students, either those in the schools left behind or those in the destination schools. However, if the magnitude of the peer effects were greater for students with low socioeconomic status, then the movement of such students into schools with more affluent peers could potentially increase overall achievement. Alternatively, however, if the students who left the public schools were the more able and more motivated students, their gains in achievement could be more than offset by the loss in achievement of the students in the schools left behind, thereby reducing overall achievement.

Such asymmetry in peer effects is quite plausible. Students whose internal motivation to learn is reinforced by an educationally rich home environment, as is true for many students with high socioeconomic status, are likely to do relatively well in most academic settings. In contrast, the performance of students from more educationally impoverished backgrounds could depend more heavily on the school. The presence of unmotivated fellow students and other associated features such as low expectations in the classroom, poor teachers and limited resources may well take its toll in significantly lower learning for such students relative to how they would perform in a school with stronger peers and higher levels of expectations, teacher quality and resources.

However, evidence on whether peer effects are asymmetric in this way is limited and inconclusive. Focusing on the spillovers from having high-ability peers, Hanushek et al. (2001) find that peer effects are positive throughout the range of student test scores, and they find no evidence of nonlinear effects. Using gender and race to define peer groups, Hoxby (2001) finds at most only limited evidence of nonlinearities. Evidence related to a more general form of peer effects that include neighborhood effects emerges from recent analysis of an experimental program of the Department of Housing and Urban Development in which families were moved out of high-poverty areas into more economically mixed neighborhoods. That analysis provides evidence that children in elementary schools who had access to schools with higher average test scores and more
affluent neighborhoods achieved at higher levels than children in the control group. It provides no support, however, for the presence of a nonlinear relationship (Ludwig, Ladd and Duncan, 2001).

On the other side, consistent with an asymmetric effect are findings that school-level measures of socioeconomic status have stronger effects on the performance of black than on white students (Coleman et al., 1966) and that racial isolation had negative impacts on student performance in North Carolina schools (Mickelson, 2001). This lack of clarity about how peer effects differ among groups rules out any clear predictions about whether a voucher program would be likely to increase or decrease the overall productivity of the education system through the mechanism of peer effects.

The Competition Effect on School Productivity

Even with a large, unrestricted voucher program, the majority of students are likely to remain in traditional public schools. In Chile’s universal voucher program, for example, the percentage of students in private schools was still below 50 percent more than ten years into the program. Hence, crucial to the argument that a universal voucher program would increase overall student achievement is that voucher-induced competition from private schools would pressure traditional public schools to become more productive and force the weaker schools to close.

There are reasons to question these predictions. One possibility emphasized by McMillan (1999) is that the greater availability of private schools may reduce parental involvement in the public schools and thereby reduce one important positive contributor to student achievement in those schools. Moreover, in contrast to the private schools, which typically have significant leeway to select their students and, therefore, to offer focused, coherent programs designed to meet the needs of those students, competition may force the public schools to offer a diverse and unfocused education program as they struggle to be attractive to all comers (Fiske and Ladd, 2000). Finally, the notion that the unproductive public schools will go out of business and that new and more effective public schools will replace them is far easier imagined than done.

The strongest claims that voucher programs have succeeded in making public schools more productive are based on two empirical studies: Jay Greene’s study of the Florida voucher program and Caroline Hoxby’s analysis of fourth-grade achievement in the Milwaukee voucher experiment. However, in both cases there are good reasons to question whether the point has been proven.

In a highly publicized study for the Manhattan Institute, Greene (2001) studied the effects of Florida’s voucher program on achievement in the public schools. Under the Florida system, schools are given grades of A, B, C, D or F. If a school receives two Fs within four years, its students are eligible for vouchers to attend private schools. Greene reports that after the first year of the program, the schools that had one F, and hence were subject to the threat of a voucher, raised their achievement significantly more than comparable schools not subject to such a threat and that their greater gains remained even after he adjusted for the statistical problem of regression to the mean.
However, with Florida data alone, Greene is unable rule out an alternative and, in my view, more convincing explanation for his findings, namely, that the improvement in the state’s low-performing schools was a response to the state’s grading of schools, rather than to the small voucher component of that program. Support for this alternative interpretation emerges from other states such as North Carolina, which, like Florida, rank schools, but, unlike Florida, do not have voucher programs. The patterns of gains in student performance in the low-ranked schools in North Carolina after the first year of that state’s program, for example, were almost identical to those found in Florida (Ladd and Glennie, 2001). The comparison between Florida and other states such as North Carolina strongly suggests that the increased scrutiny, shame and additional assistance associated with being labeled a low-performing or “failing” school is a more likely explanation for the improvement of the bottom schools in Florida than the threat of a voucher.

Hoxby (2001) carried out an analysis of fourth-grade achievement in Milwaukee public schools before and after the expansion of that city’s voucher program in 1998. She reports that the annual increase in student achievement was higher in the Milwaukee public schools most subject to competition from private schools (those with 75 percent or more of their students eligible for vouchers) than in other Milwaukee schools and even higher than in a control group of twelve other Wisconsin elementary schools. In math, for example, the annual increase in test scores of 7.1 percentile points in the schools subject to competition exceeded the increases of 5.3 and 3.7 percentile points in the other two groups, and similar patterns emerged for science and language.

However, Hoxby’s (2001) interpretation overstates the potential gains from a voucher program because she was unable to control for the changing mix of students in her treatment and control groups. The Milwaukee program was limited to low-income families. Detailed analysis of an earlier version of the Milwaukee’s means-tested voucher program indicated that the average test scores of the voucher applicants were well below those of other students in the Milwaukee system (Witte, 1999, chapter 4). Provided that pattern continued with the 1998 expansion of the Milwaukee program, we would expect to see a movement of relatively low performing students out of the treatment group of schools and a potentially large corresponding increase in the average achievement of students remaining within those schools even with no change in the productivity of the public schools.

In addition, it appears that the mix of students in Hoxby’s (2001) twelve control schools from the rest of the state may have been changing as well. My own crude analysis based on data from the National Center for Education Statistics at the district level shows that the percentages of low-income students increased on average during the 1997-2000 period in the four districts in which most of the 12 control schools were apparently located at the same time that the percentage declined by 4.7 percentage points in Milwaukee. Those increases in the proportions of challenging-to-educate children thus provide an alternative explanation for the sluggish growth in average achievement in Hoxby’s control schools relative to the schools subject to competition in Milwaukee.
Thus, it would be premature to conclude, based on the Greene (2001) and Hoxby (2001) studies, that voucher programs have unleashed strong positive impacts on the public schools. In the absence of longitudinal data for individual students that would allow the researcher to isolate impacts on achievement of the same group of students over time, any conclusions about competitive impacts are highly suspect.

Other potentially relevant sources of evidence on this question are studies of how competition from private schools and charters have affected the traditional public schools in the United States and studies of how the voucher-induced private schools have affected public schools in Chile.

The U.S. evidence from private schools provides evidence of at most small positive impacts of private schools’ competition on academic achievement in the public schools. In a comprehensive review, Belfield and Levin (2001) report that well over half of the 94 estimates in 14 studies were statistically insignificant and that any positive impacts were either substantively small or subject to question based on subsequent studies. A handful of estimates, including those by McMillan (1999), who incorporated parental involvement, suggest that competition from private schools may have a negative impact on public schools.

With respect to the effects of competition from charter schools, evidence is only now beginning to emerge in states such as California, Arizona and Michigan, where charter schools are now common. Anecdotal and interview data suggest that some school districts or schools have been responding to the establishment of charter schools in positive ways (Rofes, 1998; Gresham et al., 2000; Hess, Maranto and Milliman, 2000). For example, some school districts have set up after-school or all-day kindergarten programs, established new magnet schools, changed curriculum, empowered teachers or changed principals. In addition, some principals appear to have promoted experimentation in teaching or pursued other forms of behavior that could be viewed as positive. However, many school districts have not responded at all to the new charter schools.

Only a few studies examine impacts of charter schools on outcome measures such as achievement. Hoxby (2001) finds small positive impacts, but her estimates in this study are subject to the same sorting bias that emerges in her study of vouchers in Milwaukee. I find more methodologically persuasive a study of Michigan schools by Bettinger (1999) because of the attention it pays to various statistical problems—including the possibility that the location of charter schools may be influenced by the performance of the public schools. To deal with this simultaneity problem, the author took advantage of the role of public universities in founding charter schools to develop an exogenous instrumental variable. The study relies on data only through 1996, but it finds no impact of charter schools on public school performance.

Potentially more reliable information emerges from Chile, which has 20 years of experience with a large-scale voucher program. If competition increases achievement, one would expect the public schools in Chilean municipalities with large increases in private enrollment shares to exhibit greater gains in achievement than those subject to
less competition from the private schools. McEwan’s (2000a) study of Chilean schools uses panel data and a “differences-in-differences” methodology to sort out the effects. This statistical strategy involves comparing changes in test scores in one period to changes in test scores in the previous period. The advantage of this approach is that it controls both for unobserved determinants of achievement that are constant over time for individual schools and for unobserved time trends in each school’s achievement. McEwan’s conclusions are mixed. His preferred estimates suggest that 15 years of competition led to modest gains in achievement of about 0.16 to 0.2 standard deviations among some public schools in Santiago, Chile’s capital, but small negative effects in the rest of the country, which is home to three-quarters of the country’s population. He concludes that the results “neither refute nor provide strong support for the view that competition will lead to improvements in the quality of public schools.”

Net Impact of the Three Mechanisms on Student Achievement

The evidence suggests that the overall effect of vouchers on student achievement is likely to be small at best. Studies from the United States and other countries provide no compelling evidence in support of the view that the private sector is generally more productive than the public sector, except possibly for a subset of African-American students, nor that there are significant gains to be had from competition. Nor is there clear evidence of the asymmetric peer effects that could affect overall productivity. Empirical support for at most a small overall effect emerges from Chile, where Hseigh and Urquiola (2001) estimated the net effect of the three mechanisms based on that country’s 20-year experience with vouchers, but found only small and statistically insignificant effects of the voucher program on student achievement.

The observation that the productivity argument for a universal voucher program is weak does not rule out other types of benefits, such as those that would accrue to families who used school vouchers to achieve a better match between their values, including their religious values, and the values of the schools their children attend. It does imply, however, that the debate about voucher programs should revolve around the desirability of benefits of that type rather than around their alleged contribution to student achievement.

Impacts of a Voucher Program on Disadvantaged Students

Many supporters of vouchers believe they will improve the welfare of educationally disadvantaged students. The question of whether a voucher program will help or harm such students turns in part on whether the program is a large-scale universal program or a smaller program targeted specifically toward disadvantaged students.

A Large-Scale Universal Voucher Program

Several theoretical models of vouchers predict that the students at the bottom of the distribution in the public school sector will end up worse off under a universal voucher program (Epple and Romano, 1998; Nechyba, 1999, 2001). The main reason the students at the bottom are harmed is peer effects. With the introduction of vouchers, the more able
or motivated students leave public schools for private schools. Their departure reduces the quality of public schools and renders students remaining in those schools worse off than they would have been without the voucher program. Note that this outcome requires only that peer effects be positive, not that they have differential effects on different groups of students. Moreover, in the Epple and Romano (1998) model, even some students who switch to private schools may end up worse off. That outcome can occur when private schools are allowed to charge more than the voucher. In that case, some students will switch to private schools primarily to avoid the decline in the quality of the public schools. Even if their achievement is higher in the private school, they could be worse off because they now must pay tuition.

Some might discount the predictions of these models because the authors explicitly assume that a voucher system would have no positive impact on the productivity of the education system. However, the review of the empirical evidence in the previous section suggests that assumption is reasonable.

The data on educational outcomes from Chile’s universal voucher program generally supports the predictions of the theoretical models. Chile’s voucher program did induce the higher-income and -ability students out of the public sector schools. Cross-sectional quantile regressions for that country, which allow one to look at the performance of different parts of the student distribution, show that voucher-induced expansions of private schools widened the variation in educational outcomes across students (Hsieh and Urquiola, 2002). New Zealand’s experience with universal public school choice was similar, although conclusions about achievement are harder to confirm because of the limited achievement data in that country. However, there is little doubt that the expansion of choice in that country exacerbated the problems of the schools at the bottom of the distribution and reduced the ability of those schools to provide an adequate education (Fiske and Ladd, 2000).

Of course, some disadvantaged students in both countries were made better-off because vouchers in Chile and public school choice in New Zealand gave those students access to some schools that previously were outside their financial reach. This observation is important and complicates any discussion of voucher programs. From an ethical perspective, it is hard to justify denying schooling options to such children because their families are poor or because their departure may reduce the quality of education of those who remain behind. Providing additional choices to such families is a desirable goal.

The adverse effects of large-scale voucher programs on the students left behind highlights the need to shape voucher policies in ways that could minimize those effects. One starting point might be to adjust the voucher amount both to the characteristics of the students (with costly-to-educate students getting significantly larger vouchers) and possibly to the characteristics of the schools (for example, students attending economically integrated schools would receive larger vouchers than those in homogeneous schools). Such schemes, however, are likely to be politically contentious and difficult to implement. A second approach is to place restrictions on the use of vouchers, such as prohibiting the participating private schools from charging additional tuition or fees above the voucher amount (a prohibition that is part of Chile’s program).
and requiring that oversubscribed private schools select students randomly (as is the case in Milwaukee and Cleveland). With such constraints, public funds would support education only in schools that were available to all students, and all students would continue to be guaranteed an education at no out-of-pocket expense, whether it be in a public or private school. These constraints would make voucher-financed private schools similar to charter schools. Indeed, one prominent supporter of charter schools argues that these two characteristics of charter schools—equal access and no fees—are what make charter schools preferable to voucher-financed private schools (Hassel, 1998).

However, even if a universal voucher program were modified in these ways, the fundamental problem facing the U.S. education system with respect to disadvantaged students would remain. That problem is the significant concentrations of difficult-to-educate students in some schools. The challenge is to find a way to expand the educational choices available to families, while at the same time reducing those concentrations. That challenge can best be met with some form of controlled or managed choice among public schools, as forcefully advocated in a recent Task Force Report from the Century Foundation (2002). Under such a system, families would specify their preferred schools. Students would then be assigned to schools based on those preferences, but with attention to the mix of students in each school, either the racial mix (as was long the case in the managed choice program in Cambridge, Massachusetts) or the economic mix as advocated by the Task Force (and recently introduced in Cambridge). Any such system of controlled choice would require that the public authorities ultimately have the power to assign students to schools. In addition, it would require targeted investments focused on teaching and learning in the schools that were not successful in attracting students so as to promote healthy competition among all schools. Since private schools are likely to be reluctant to participate in a program of that type, vouchers would not be a logical component of such a strategy.

A Means-Tested Voucher Program Limited to Low-Income Families

If the goal were to use a voucher program specifically to assist low-income families, at a minimum, the program would have to be means tested. Evidence from Milwaukee suggests that a means-tested voucher program can be successfully designed to serve low-income families, especially those whose children are unsuccessful in the public schools (Witte, 1999). The main advantage of such programs is that parents tend to be more satisfied with their new schools than with their assigned public schools. In the fifth year of the Milwaukee voucher program, for example, more than three-quarters of choice parents gave their child’s school a grade of A or B. Similarly, in the three urban privately funded voucher programs, 40 percent of the private school parents gave their schools an A compared with 14 percent of the control group (Howell and Peterson, 2002). Moreover, satisfaction levels among private schools parents were higher with respect to all the major components of the school: the academic program, school safety, parental involvement and class size (Howell and Peterson, 2002).

Of course, some of this increased satisfaction may reflect not the specific policies of the schools but rather a different, more congenial or more motivated set of peers in the new schools. To the extent that families opt for schools in which their children will have peers
with higher socioeconomic status, their behavior complicates the policy discussion, because not all families can achieve that end. Nonetheless, it is hard to argue that low-income families should be denied opportunities to benefit from such choices simply because they are poor.

Even for means-tested programs, however, design matters. For example, the Milwaukee program required that schools accept the voucher in lieu of all tuition and fees and that oversubscribed schools select students randomly—a design that helps reduce the negative impact of vouchers on disadvantaged students. In contrast, consider the design of the Children’s Scholarship Fund (CSF), a privately funded national school voucher program that provides scholarships for children from low- to moderate-income families (up to 270 percent of the poverty level) to attend private schools, with the scholarships scaled to the income of the family. Although the program reached large numbers of African-American low-income and minority households, only 30 percent of all families offered vouchers ended up using them. Forty-five percent of the decliners said that they could not afford their preferred school, 10 percent said no space was available, and 8 percent cited transportation problems (Campbell, West and Peterson, 2001). Thus, when tuitions are not limited to the amount of the voucher and transportation is not provided, many families will not be able to benefit from the voucher program.

Additional concerns arise from the high rates of attrition from voucher programs over time. In the early years of the Milwaukee program, dropout rates were very high: 54 percent of the voucher recipients did not return to their private school after the first year of the program (Witte, 1999). The high attrition rate in that program might be explained in part by the poor quality of some of the participating private schools, given that religious schools were not permitted to participate and the fact that one major private school closed. However, high dropout rates also emerged in the privately funded New York City voucher program. By the end of the third year of that program, 38 percent of the voucher users had left their voucher-subsidized private school (Howell, 2002). Multivariate analyses of those who were offered vouchers, used them and then dropped out indicate that while African-Americans were more likely than other racial groups to accept an offer of a voucher, they were also far more likely to drop out each year (Howell, 2002). This finding raises additional questions about the ability of even a means-tested voucher program to meet the needs of low-income African-American students.

Given the serious educational challenges facing disadvantaged students, particularly those living in areas of concentrated poverty, it is hard to argue against any policy, including means-tested vouchers, that might improve the educational experiences of some students. In some situations, means-tested vouchers could play a useful role in providing additional options to students whose schooling choices would otherwise be severely constrained. Even such voucher programs, however, would need to be embedded in a larger strategy of education reform that focused on teaching and learning in the public schools and one that provided greater choice within the public school system.
CONCLUSION

Contrary to the claims of many voucher advocates, widespread use of school vouchers is not likely to generate substantial gains in the productivity of the U.S. K-12 education system. Any gains in overall student achievement are likely to be small at best. Moreover, given the tendency of parents to judge schools in part by the characteristics of the students in the school, a universal voucher system would undoubtedly harm large numbers of disadvantaged students. Although small means-tested voucher programs might provide a helpful safety valve for some children, policymakers should be under no illusion that such programs will address the fundamental challenge of providing an adequate education to the large numbers of disadvantaged students in many of our large cities. At the same time, there are good arguments for giving families, especially those who are economically disadvantaged, more power to choose the schools their children attend. The challenge for policymakers is to find ways to expand parental choices without excessively privileging the interests of individual families over the social interests that justify the public funding of K-12 education.

§ 8.3. The Politics of Vouchers


INTRODUCTION

[In the conventional historical account, school is identified with the political right:] free-market economist Milton Friedman, attempts to defy Brown v. Board of Education, wealthy conservative philanthropists, and the attacks on the public school bureaucracy by Ronald Reagan and George W. Bush.

It turns out that the conventional history is incomplete. This is not to say that it is wrong, however. It is true, for example, that southern states used choice to evade Brown. Some adopted “freedom of choice” plans, which purported to give choice to black and white families, but in fact were an attempt to keep black and white students in the same segregated schools they attended before Brown. Others gave tuition grants to white students that allowed them to attend private, segregated white academies. Though these plans were eventually struck down, they effectively delayed Brown’s implementation by at least a decade. So it is correct to conclude . . . that “choice has a history of unlawfully segregating students.”

But too often missing from the historical account is the left’s substantial—indeed, I would say leading—contribution to the development of school choice. . . . [C]hoice has deep roots in liberal educational reform movements, the civil rights movement, and black nationalism. While progressives have not focused on choice exclusively, . . . the left has,
at important junctures, seen a role for choice in promoting educational opportunity for poor and minority children. . . . [U]nderstanding the history of progressive choice proposals—including even school vouchers—might offer today’s liberals a way to have a more nuanced conversation about school choice. . . .

. . . In Part I, I outline how sheer necessity forced recently freed slave communities to create schools for themselves and their children during Reconstruction. In Part II, I examine the alternative schools created during the civil rights movement of the 1960s. My emphasis here is on the Mississippi Freedom Summer of 1964, in which civil rights organizations and volunteer teachers from the North responded to the inadequacies of the Mississippi education system by establishing freedom schools for blacks.

Though most of the freedom schools disbanded at the end of the summer of 1964, their legacy would influence the next chapter of the progressive history of school choice—the northern free schools of the late 1960s and early 1970s. Free schools, the subject of Part III, were alternative, independent, and privately funded. They were started by progressive educators, including many former public school teachers and some veterans of Freedom Summer, who had come to believe that they could reform public education only by working outside of the system.

The late 1960s also saw the public school bureaucracy challenged by the community control movement, the topic of Part IV. Community control advocates included civil rights organizations, black nationalists, and some members of the liberal political establishment. They demanded that ghetto residents have more control over their neighborhood schools. While not itself a school choice initiative, the community control movement is an important part of the narrative. Like some choice proposals, it was premised on the notion that the public system was unwilling or unable to meet the needs of poor and working-class black children. Community control supporters also shared choice advocates’ belief that taking control from the bureaucracy and giving it to community members was an important part of the solution.

As I will address in Part V, these views found their ultimate expression in the claim that the state should provide vouchers that parents could use to pay for schools of their choice. Despite the voucher movement’s current association with conservatives, I will explore how vouchers became part of the political landscape through the advocacy of influential liberal educational advocates in the late 1960s. These progressives outlined a vision of school vouchers that was clearly motivated by integrationist and equity concerns—the same impulses that had led to Brown and Bolling [v. Sharpe].

Finally, in Part VI, I will sketch out some of the implications of this history. In particular, I will argue that a close attention to the history of school choice suggests that while today’s voucher proposals will not likely further a robust version of the Brown/Bolling promise of racial justice, it is possible to imagine a progressive vision of vouchers that would. . . .
If the progressive contribution to school choice generally has been too often overlooked, nowhere is this more true than in the debate over school vouchers. Missing from many accounts is the fact that, during the late 1960s, some leading progressives outlined voucher plans. Indeed, though proposed by Milton Friedman in 1955 (and mentioned by Adam Smith centuries before that), vouchers—the idea that the state would give families a sum of money that they could use to enroll their child at the public or private school of their choice—had been largely ignored until its revival by 1960s liberals.

Those progressives who endorsed vouchers did so for many of the same reasons that had led others to advocate for freedom schools, free schools, and community control. Principal among them was anger at how inner-city schools had failed black children. Writing in the *New York Times Magazine* in 1968, during the height of the community control debates, Christopher Jencks argued: “The origin of the crisis is simple. The public schools have not been able to teach most black children to read and write or to add and subtract competently. This is not the children’s fault.” For Jencks, “[g]hetto schools have therefore become little more than custodial institutions for keeping the children off the street.” As had some of the free schoolers, voucher advocates argued that the system was not just inept but actually harmful and that liberals should no longer support it. Ted Sizer and Phillip Whitten, in defending their voucher proposal, wrote, “[t]hose who would argue that our proposal would destroy the public schools raise a false issue. A system of public schools which destroys rather than develops human potential now exists. It is not in the public interest . . . . [I]t does not deserve to survive.”

Progressive voucher proponents were aware of, and largely sympathetic to, movements such as community control and free schools. Jencks and Sizer, for example, saw vouchers as a way to achieve some of the same goals. Like community control, vouchers would “distribute power, and give parents some options, some clear leverage on their neighborhood school.” Indeed, Jencks suggested that the “community solidarity and pride in the ghetto” that would come from independent black schools were among the strongest arguments for vouchers. Moreover, vouchers would be a way to provide a revenue stream for some of the free schools, which, as I have mentioned, often closed for lack of funding. As Sizer envisioned it, vouchers would allow an area to have, “cheek by jowl, several public schools—for example, one run by the city (PS 121), one by the Catholic Church (St. Mary’s School), and one by the black community (the Martin Luther King Freedom School).”

These early voucher plans were self-consciously designed to maximize equity and racial justice. The advocates placed faith in the market as against the alternatives, but they acknowledged (unlike Friedman and many of today’s voucher proponents) that the private market often did not serve the poor especially well. Their programs were designed to maximize the likelihood that poor children and families would succeed in the market. Because “a child from a severely restricted background will require more expensive services than one from a wealthier family,” Sizer offered a plan in which “the poorer the child, the more valuable the voucher.” Recognizing the importance of giving inner-city children the chance to attend suburban schools, Sizer argued that the voucher amounts
must be great enough that suburban districts would “be bribed into taking many poor youngsters.” Moreover, Sizer and Whitten insisted that education reform in and of itself would not be enough to achieve genuine equality. They proposed a “Poor Children’s Bill of Rights,” and suggested that vouchers “must be part of a package, one which surely must include some form of guaranteed annual income and the provision of health and welfare services at a level of accommodation far higher than at present.”

With these equity goals in mind, the Johnson administration’s Office of Economic Opportunity gave a grant to Jencks and his colleagues at Harvard’s Center for the Study of Public Policy to design an equity-oriented voucher system. The Jencks plan was explicit in its motivation: “a program which seems to improve education must therefore focus on inequality, attempting to close the gap between the advantaged and disadvantaged.” Closing this gap required disadvantaged children to get more than equal educational resources—they needed the smallest classes and the best teachers. Moreover, because a student’s classmates are among her greatest resources, enrollment patterns should be altered so that disadvantaged children would have the chance to attend school with more advantaged classmates. Accordingly, the Jencks plan envisioned that both public and private schools would participate in the plan. Like Sizer, Jencks’s team hoped that race and class integration could be achieved by bonus vouchers—up to double the basic voucher—that would make poor children attractive to successful schools.

Of special concern to the Jencks team was the prospect that schools could selectively admit students. They pointed out that a means-tested voucher program did not by itself ensure that schools would not discriminate against certain categories of low-income students.

Incentives which reward the admission of low-income applicants will initially result in schools’ seeking out families which are short on cash but long on other “desirable” characteristics, such as literacy, initiative, and self-discipline. Unless some machinery is established for preventing discrimination on the basis of IQ and behavior patterns, overapplied schools will get big bonuses for taking the most easily educated children of poor families, while leaving the others to underapplied schools.

In response, the Jencks team proposed that participating schools could use traditional admissions criteria for up to half the entering class, but the other half had to be accepted by lottery.

Like the other reforms I have discussed, the progressive voucher movement was relatively short-lived. A modified version of the Jencks plan was ultimately tested in Alum Rock, California, a mixed-race, mixed-income district within the city limits of San Jose. The Alum Rock experiment went forward for five years, with inconclusive results, and the Nixon administration did not continue the grant. The political constituency for the project didn’t materialize, as many liberals did not support Jencks’ proposal, despite its progressive heritage. Even some who may have been sympathetic to the Jencks plan as crafted ended up opposing it. They believed it would never be enacted without substantial modifications that would make it less palatable. For example, according to American Federation of Teachers president Albert Shanker:
There is no legislature in the nation that will enact a voucher plan that will give substantially more money to each poor child and bar private schools from freely selecting the students they want and rejecting those they don’t want. By the time Jencks’ model goes through the political-legislative wringer, it will become the conservative model that its author himself opposes.

Nor did conservatives back the plan, for they wanted an even purer free-market approach with fewer regulations than accompanied the Jencks model. Without support from any constituency, the Jencks voucher plan died.

VI. REFLECTIONS ON TODAY’S SCHOOL CHOICE DEBATES

. . . I have tried to suggest that the history of school choice is substantially more complicated than we have traditionally understood. Progressives have proposed a variety of school choice schemes, and this lesser-known heritage is at least as important as the Friedman-inspired deregulation impulses or the conservative anti-desegregation movement. In this final Part, I would like to offer some thoughts about how thinking about these progressive school choice movements might influence today’s conversation about school choice.

One of the most remarkable features of current choice debates is the faith that progressive educational reformers place in the existing public school system. Today, it is common to hear arguments that school choice proposals should be rejected because they do not “do anything for the public school system, and that’s where our folks are being educated.” Before experimenting with choice, it is argued, we must “fix the public schools.” My review of the history of school choice reveals the novelty of this position. That doesn’t mean it is wrong, of course. Yet it is noteworthy that, in contrast to today’s faith in reform from within, many liberals in the 1960s believed that only developing progressive alternatives outside the public system would reform it.

What explains this reversal—how did advocates for racial justice come to defend the school bureaucracy they once attacked? I suspect that it is part of a larger trend in which the left has abandoned some of the anti-bureaucratic ideology that marked its politics during the 1960s. Moreover, the politics of race is likely at play here as well. Many of the urban public school administrations that the left once attacked as white, middle-class enclaves now are the province of middle-class black managers. In his study of urban school systems, Jeffrey Henig found that in Baltimore, Detroit and Washington, D.C., the public school systems are the city’s largest single employer. Many in the black community, including much of the civil rights leadership, have been less likely to criticize the under-performance of these black-run systems.

But the most interesting implications of this look back at the history of progressive choice proposals concerns the conversation about school vouchers, the most controversial of today’s choice proposals. Today, many conservatives not only enthusiastically push vouchers, they wrap their initiatives in the mantle of civil rights. A few days after the Supreme Court upheld Cleveland, Ohio’s voucher plan, President Bush hailed the decision at a “Rally on Inner City Compassion” in downtown Cleveland. In Brown v.
Board of Education, said Bush, the Court “declared that our nation cannot have two educational systems, and that was the right decision.” The voucher ruling was similar, he said. “Last week, what’s notable and important is that the court declared that our nation will not accept one education system for those who can afford to send their children to a school of their choice and for those who can’t, and that’s just as historic.” Nor was Bush alone in comparing Brown and Zelman. Secretary of Education Rod Paige, for example, argued:

Brown v. Board of Education changed American education forever. I know because I grew up in the South when schools were segregated. With Brown, education became a civil rights issue, and the decision introduced a civil rights revolution that continues to this day. Zelman v. Simmons-Harris holds the same potential. It recasts the education debates in this country, encouraging a new civil rights revolution and ushering in a “new birth of freedom” for parents and their children everywhere in America.

Despite such rhetoric, an examination of the Jencks and Sizer proposals reveal the limitations of today’s voucher plans. Consider first the voucher amount. Whereas the Jencks plan envisioned a base voucher that was equal to the amount spent per pupil by the public schools in the region, today’s voucher plans typically are much stingier. The Cleveland plan, for example, provides vouchers of $2,250—by contrast, school districts receive over $7,000 per pupil. Moreover, because the purpose of the Jencks plan was to reallocate resources toward the poor, low-income students would qualify for bonus vouchers up to double the base voucher. To again use Cleveland, depending on her level of disadvantage, a poor student under the Jencks plan today might qualify for up to a $14,000 voucher, rather than the $2,250 voucher in effect in Cleveland today.

This raises the question of the participation of suburban schools in voucher plans. Race and class integration were central to Sizer’s and Jencks’ vision. They believed that their equity-oriented voucher plans could achieve integration by making disadvantaged students attractive to middle-class schools. Today’s voucher plans, by contrast, largely abandon this goal. Instead, they assume that inner-city students will continue to go to school with others of their same race and class background. The only change is to the type of school, not the composition of the student body.

In this respect, today’s voucher plans have less in common with Brown or Bolling than they do with the compromise President Richard Nixon proposed as part of his opposition to busing. Nixon argued that rather than bus students to integrated schools, inner-city children should instead get extra resources and stay in their own schools. Nixon was explicit about this in his nationally televised address explaining that he would oppose busing but support additional funding for inner-city schools. Nixon said, “[i]t is time for us to make a national commitment to see that the schools in the central cities are upgraded so that the children who go there will have just as good a chance to get quality education as do the children who go to school in the suburbs.” The Supreme Court reached a similar conclusion when it struck down a metropolitan-wide busing plan in Detroit, and upheld a remedy that included compensatory education programs for inner-
city students (including remedial reading, improved teacher training, bias-free testing, and guidance and counseling programs.)

Substitute “private schools” for “extra resources” and the link between school vouchers and Nixon/Milliken becomes apparent. Without the participation of suburban schools, inner-city private schools become simply another attempt to fix the urban school crisis without altering the race and class segregation levels of those schools. As Robert Vischer has written: “[E]specially in lower-income, minority neighborhoods, the only non-public school options [available to voucher recipients] may very well be church-affiliated schools, which will likely reflect the skewed racial compositions of their congregations and neighborhoods.” The limited existing data suggests that current voucher programs either have no impact on integration or only a slightly positive integrative effect. Given the overwhelming evidence that the socioeconomic background of a student’s peer group is among the most important determinants of his academic achievement, one cannot be too optimistic about the likelihood of systemic success for a reform that ignores class integration.

Indeed, there is an important aspect in which today’s voucher plans may serve to reinforce segregated school patterns in much the same way as did the South’s post-\textit{Brown} tuition grant programs. Those programs, it is worth recalling, were not explicitly racially discriminatory. Typical was Virginia, which made tuition grants available to “[e]very child in the Commonwealth . . . who desires to attend a nonsectarian private school.” Despite their facial neutrality, the programs had the effect of furthering segregation, because the private schools that were available to voucher recipients were all-white and had the freedom to discriminate in admissions. Against this background, what effect will today’s voucher programs have on patterns of school segregation? While current voucher programs prohibit schools from using racially discriminatory admissions criteria, existing patterns of racial segregation will leave most parents with private schooling options that are de facto one race schools.

Of course, one response is that the private schools are no more racially segregated than the public schools to which a neighborhood’s children already had access. In that sense, the voucher plans do not make segregation any worse than it already is. There is some truth here—when opponents claim that vouchers are a scheme to re-segregate the schools they seem to overlook the extraordinarily high segregation levels already existing in public schools. Moreover, it is beyond question that segregation cannot be the only benchmark. Some schools are able to achieve success against the odds, and whether those schools are voucher schools, charter schools, or other alternatives, we need more of them, even if they remain segregated by race or class. At the same time, the response that vouchers (or any other reform) will not increase segregation levels hardly settles the matter. Instead it invites us to ask: Why should we settle for voucher initiatives that fall so short of the ambitious equity and integrationist goals of Sizer and Jencks?

While conservatives have pushed voucher programs that fall short, progressives—especially leading liberal advocacy groups—generally have denounced the very idea of vouchers. I believe this has been the greatest cost of overlooking the history of progressive voucher efforts. Because progressives have by and large focused on the evils
of vouchers, insufficient thought has gone into whether a modern-day progressive vision of vouchers is possible. Instead of asking whether vouchers are good or bad, progressives might do well to consider that vouchers are neither and both—it all depends on how the plan is constructed.

What if voucher plans today were designed to both re-allocate additional educational resources to disadvantaged students and to increase class integration? What if they were animated by the integrationist impulses of *Brown* and *Bolling*? They would look more like those proposed by Sizer and Jencks. They would need to be more generous than today’s plans and designed with the goal that some low-income students would have access to economically integrated schools. Given the extent of residential segregation in our nation’s metropolitan areas, in many regions this would require the participation of some suburban school districts.

There is a moral claim that progressives might make on behalf of economically integrated schools. One of the ironies of today’s voucher debate is that the moral passion that once animated advocates of free schools, community control, and progressive vouchers is now marshaled on behalf of voucher plans that modern progressives largely oppose. Joseph Viteritti, for example, argues that vouchers for private schools “could help level the playing field by offering poor children opportunities more akin to those enjoyed by the middle class.” Viteritti’s point is a powerful one, but its implications are more profound than he acknowledges. After all, for most middle-class parents in urban areas, choice means quality suburban schools, not private schools. So if the point is to offer poor children “opportunities more akin to those enjoyed by the middle class,” access to middle-class suburban schools would seem more important than access to inner-city parochial schools. Returning to the moral dimension, if it is unjust and unfair to deny a poor child in Washington, D.C. the chance to attend a private school because his parents cannot afford the tuition, why isn’t it equally so to deny him the chance to attend a suburban school because his parents can’t afford to buy a house there? That question was central to the thinking of the progressive voucher movement of the late 1960s and early 1970s, but is largely ignored today.

While some have pointed out the importance of broadening voucher initiatives along these lines, too many progressives have ignored the issue. That is a mistake. Given the success of voucher advocates at keeping vouchers at the center of the national education policy debate, the absence of a progressive vision of vouchers has consequences. But, while some voucher plans can be dangerous, as Jencks and Sizer showed thirty years ago, a properly constructed one has the potential to increase educational opportunities for disadvantaged children in a way that should appeal to all—including progressives.

**INTRODUCTION**

In *Zelman v. Simmons-Harris*, the U.S. Supreme Court upheld the constitutionality of a Cleveland program that provided school vouchers to low-income parents seeking private school alternatives for their children. Under Cleveland’s voucher plan, parents could theoretically use the voucher at religious schools, secular private schools, or suburban public schools. But few secular private schools and no suburban public schools chose to join the program. So, for most parents, the voucher option was a religious private school. Voucher opponents used this fact to argue that the plan amounted to state funding of religion and thus violated the First Amendment’s Establishment Clause. *Zelman* rejected this challenge, and the opinion was widely heralded as of great historical significance. Clint Bolick, the lawyer who led the defense of school voucher programs, called *Zelman* “one of the most important constitutional cases ever.” Bolick’s analysis was shared by more objective observers as well. John Jeffries and Jim Ryan, for example, said that *Zelman* addressed “the most important church-state issue of our time: whether publicly funded vouchers may be used at private, religious schools without violating the Establishment Clause.”

*Zelman* did not directly affect many students. In 2002, when the case was decided, voucher plans existed in only three states, serving fewer than 20,000 students. *Zelman* was thought to be important because many assumed that once the Court held vouchers to be constitutional, states would rush to implement such plans. For many, the uncertain legality of school vouchers had been a reason not to institute voucher programs. With no remaining obstacles under federal law, voucher proponents could fairly predict that “[s]chool choice is an inevitability.”

Yet, in the years since *Zelman*, school vouchers have made little political headway. They have been proposed in a variety of cities and states, but have overwhelmingly been rejected. This is just as true in states run by Republicans as in those led by Democrats. Since *Zelman*, only three jurisdictions have adopted voucher plans, while proposals have failed in over thirty-four states.

It is tempting to think that vouchers simply need more time to gain popularity—after all, *Zelman* was only decided five years ago. While that is surely part of the explanation for the slow growth of vouchers, it is not the only one. We can see this by comparing the relatively slow pace of voucher expansion to the rapid growth of two other forms of school choice that exploded in the preceding decades. Like vouchers, home schooling and charter schools were new innovations at one point. But they both grew much faster. In 1983 only four states had laws that explicitly permitted home schooling; ten years later all fifty states had such laws. Charter schools have had a similarly rapid rise. The nation’s first charter schools were authorized in Minnesota in 1991; by the 2004-05 school year,
there were approximately 4000 charter schools in forty states and the District of Columbia.

Why have school vouchers failed to garner the support that so many assumed would follow the Court’s decision in Zelman? To answer this question requires looking back at the evolution of the school voucher movement. In this Article, I suggest that the story is one of religion, race, and politics. It is a story of religion because religious conservatives—especially Christian conservatives—once championed school vouchers and other forms of private school choice as their leading education priority. Christian conservatives were drawn to vouchers because they sought schools that would reinforce their religious beliefs and values—what I call the “values claim” for vouchers. The values claim was central for most of the history of the voucher movement.

This is a story about race, too, for the leaders of the voucher movement made a disciplined effort to define school vouchers as part of the struggle for racial justice and educational opportunity. In so doing, they developed an alternative rationale for school vouchers—what I call the “racial-justice claim”—which emphasized the right of low-income and minority parents to send their children to academically rigorous private schools. The racial-justice claim had political and legal advantages for the voucher movement. It attracted an additional constituency—black parents—and made voucher plans less vulnerable to Establishment Clause challenges. As I will explain, however, jettisoning the values claim that appealed to religious conservatives may have weakened the post-Zelman political movement for vouchers.

This is also a story about politics, for by the time Zelman was decided, a political consensus had emerged emphasizing the importance of “accountability” in education. Best captured by the federal No Child Left Behind legislation, the new politics of accountability increased state and federal oversight over individual schools and districts. For schools, it means less local control, more tests, stricter standards, and various other regulations imposed by governing authorities. Although No Child Left Behind does not govern private schools receiving vouchers, there is growing pressure for increased government oversight of those schools. This threat of governmental regulation is anathema to conservative Christian educators, driving them further away from a school voucher movement about which they were already increasingly ambivalent.

In Part I of this Article, I outline the evolution of the values claim for school vouchers. In the early twentieth century, the idea of giving parents public money to pay religious school tuition was advocated mainly by Catholics, who had long battled with Protestants over whose religion and values would be taught in school. Catholics lost this battle, and when they did, many left the public system. They built their own schools and sought state funding for them. State funding of private schooling had not initially appealed to Protestants but gained in popularity among a wider swath of religious voters in the 1970s and 1980s. The principal rationale was that the public schools were becoming increasingly secular and hostile to religion. Over time, courts had prohibited school prayer and the teaching of creationism. They also rejected a variety of challenges by religious parents to the secularism of school textbooks. In response to these defeats, evangelical Christians and others began to argue that parents should send their children to
schools that reinforced, or at least respected, their core beliefs and values. Moreover, like the Catholics before them, they began to argue that those who made such a choice should receive some sort of government support. Otherwise, some would not be able to afford religious schools, and even those who could afford them would be forced to pay twice—once in taxes for the public schools they did not use, and again for the religious schools that they did.

As I describe in Part II, during the 1990s the values claim took a back seat as voucher advocates promoted a racial-justice claim in its place. This racial-justice claim came to define the litigation strategy in defense of vouchers, and, I argue, was essential to achieving the legal victory in Zelman. The racial-justice claim asserted that vouchers provided educational emancipation for poor students, mostly black, trapped in dysfunctional urban districts. Vouchers were hailed as a way for these students to gain access to schools in which they could acquire the academic skills they needed to succeed in college and the workforce. This rationale for vouchers was a significant departure from the earlier theory that had attracted religious conservatives. Unlike the values rationale, the racial-justice claim did not assert that public schools were teaching the wrong values to children or contradicting the parents’ religion. Rather, it argued that the public schools were not teaching children the necessary academic skills, and that private schools could do better. Unlike the values rationale, the racial-justice claim did not defend the rights of all families whose values conflicted with school authorities. Under this claim, the victims were a smaller group of low-income, inner-city children.

In Part III, I explore the implications of replacing the values claim with the racial-justice claim. On the one hand, the new voucher movement is more attractive to a contingent of African American parents and some of those sympathetic to their plight. On the other hand, it holds less appeal for religious conservatives. Religious conservatives had not, by and large, objected to schools on the ground that they were insufficiently academically rigorous, and they are less attracted to a movement that defines the problem in this way. Furthermore, the new voucher movement, wrapped in the mantle of racial justice, promotes school-choice programs that are targeted at low-income students in urban districts. White religious conservatives do not generally live in these urban districts and do not stand to benefit from the programs.

But even if voucher programs targeted at low-income urban districts did not appeal to religious conservatives, perhaps these limited plans would provide voucher proponents with a toehold that would position them to enact more expansive voucher plans. This was certainly the stated strategy of some voucher proponents. But I suggest a reason to question the effectiveness of this approach. I argue that the new voucher movement will have trouble attracting religious conservatives because of the rise of the accountability movement in education and its impact on voucher programs. The original movement for private school choice was grounded in the notion, shared by libertarians and religious conservatives, that private schools should be largely free of government regulation. Zelman, however, gave the green light to the new voucher movement at exactly the same time that state and national education policy had come to demand greater oversight of all schools, including private schools accepting vouchers. This accountability involves increased regulation of individual schools and demands that they provide information to
various governmental authorities. This sort of regulation is opposed by religious groups
generally and evangelical educators specifically. As a result, modern voucher programs
are replete with government strings that many religious conservatives reject. This
combination of circumstances has led to a modern voucher movement that has received
constitutional approval, but may lack the necessary political support to thrive. Thus, I
predict that Zelman will end up mattering much less than many had thought it would.

In addition to being historical and predictive, the analysis advanced in this Article is
normative. I believe that the racial-justice claim for school vouchers has been grossly
overstated by voucher advocates. But I nonetheless argue that it would be unfortunate if
my prediction about the demise of vouchers proves accurate. While a central theme of
this Article is that school vouchers serve the interests of groups other than blacks, I also
argue that there is limited (and inconclusive) evidence that they improve educational
outcomes for low-income African American children. Voucher plans should continue at
least until they can be more thoroughly evaluated to determine their impact on a group so
in need of better educational opportunities. . . .

CONCLUSION

In this Article, I have explored the interplay between religion, race, and politics in the
movement for school vouchers. In investigating why vouchers have not met with more
political success following the Supreme Court’s decision in Zelman, I have emphasized
two factors—the choice to reframe school vouchers in racial-justice terms and the new
politics of accountability—that have combined to limit the appeal of vouchers to religious
conservatives, an important constituency. Further, I have argued that the demise of
vouchers would be unfortunate, since the reform deserves a life long enough to see
whether it can fulfill—even partially—the racial-justice promises of its advocates.

This is not to say that vouchers are dead. The accountability movement that has indirectly
undermined support for voucher programs has an uncertain future. NCLB is under attack
on a number of fronts. Many educators argue that it demoralizes teachers, encourages
poor instruction, and demeans students and school professionals. Some conservatives
claim that it is an unwarranted federal intrusion into education, a sphere best left to states
and local communities. If these critics prevail, the accountability era in education may
end up being short-lived, and religious conservatives will be less fearful of regulation of
voucher schools. Moreover, voucher advocates are resourceful and have recently adopted
a strategy that conservatives have long chastised liberals for—trying to win in the courts
what they cannot achieve in the legislature. In New Jersey, for example, voucher
proponents have filed suit arguing that the state constitutional guarantee of a “thorough
and efficient system of free public schools” means that the state should provide vouchers
to parents in low-performing districts.

But the challenges to the success of vouchers are serious. The public education
establishment, its unions, and its allies among traditional civil rights organizations remain
staunch voucher opponents. These groups have launched a variety of legal challenges to
vouchers under state constitutions, some of which have been successful. Charter schools,
which some feared would pave the way for vouchers, have in some states had the
opposite effect—they have reduced enthusiasm for vouchers by giving dissatisfied parents other options.

Further complicating the political future of the voucher movement is the inherent tension between the values and racial-justice claims, which I have discussed throughout the Article. For example, the racial-justice claim envisions restrictions on the ability of voucher schools to choose which students they want to admit. Most Americans—and especially minority and low-income voters—support such restrictions, and some current voucher plans prohibit selective admission. Yet, as I have shown, religious conservatives are equally committed to protecting the admissions autonomy of their schools, and the values claim supports such a view. These differences threaten to destroy the fragile voucher coalition. In Milwaukee, for example, African American community activist and legislator Polly Williams, one of the founders of the voucher movement, fought hard to limit the ability of schools to select students. But Catholic school administrators disagreed with her, and Williams has become alienated from the larger voucher coalition. Because minority voters and their leaders support regulations requiring lottery admission as strongly as religious conservatives are opposed to them, this dispute—and others like it—will continue to threaten the voucher movement.

In the face of this complicated political and legal environment, it is increasingly evident that vouchers will need support from additional constituencies for the reform to become widespread. But the voucher movement may have painted itself into a corner. In justifying vouchers as a means to an academically superior education, voucher advocates have tied their fortunes to research results proving that claim. The findings so far, viewed in the light most sympathetic to voucher defenders, show that religious schools have an impact on the academic performance of inner-city African Americans, but not on the performance of whites or Hispanics. To date, voucher proponents have trumpeted these findings as a justification for voucher programs targeted toward inner-city blacks.

But those same findings can and will be used by voucher opponents to argue that there is no justification for broader voucher plans aimed at whites or Hispanics. The difficulty of “selling” vouchers to white suburban voters—a group whose support is essential to any educational reform—received inadvertent public acknowledgment from President Bush. “There are a lot of Republicans who don’t like vouchers,” Bush whispered before a White House event, without realizing that his podium microphone was broadcasting his comments to the press. “They come from wealthy suburban districts who are scared to death of irritating the public school movement, and their schools are good.” The argument for vouchers as a means to quality education will require convincing some of these voters that voucher schools achieve better academic results than their current schools. This may turn out to be an impossible standard for the voucher movement to reach; to date, even the research from voucher advocates like Howell and Peterson does not support this claim.

So who will clamor for vouchers? If Christian conservatives are insufficiently moved by racial-justice claims and fearful of government regulation, and if suburban voters see no need to adopt a reform that has only succeeded with inner-city blacks, the voucher movement’s appeal may be limited to the African American constituency it cultivated on
the road to *Zelman*. That would be the ultimate irony, because thinkers as different as economist Milton Friedman on the right and law professor Derrick Bell on the left both agree on one thing: As long as a program is perceived as benefiting solely those with little power in society, it is unlikely to prosper for long.

---

§ 8.4. Vouchers in Other Fields


Most citizens would consider it shockingly unethical for an attorney representing one side in a lawsuit to be selected or paid, even indirectly, by the opposing party. Yet such principles are violated routinely in this country on a massive scale. In criminal cases, the great majority of defense attorneys are paid directly or indirectly by the prosecuting party, the state.

The reason that financial conflicts of interest are problematic for the client is obvious; the decisions of the attorney are bound to be affected by the desires of his actual employer. That is true for public defenders and assigned counsel in criminal cases, just as it is for private attorneys in civil cases. While the lawyers, and those who assign them to cases—judges, other government officials, or private firms contracting with government—are no doubt interested in preventing conviction of the innocent, they are less strongly committed to that objective than innocent defendants are. And various officials are likely to have other objectives, such as reducing court backlog, that conflict with the goal of acquitting the innocent. They also may have little commitment to one of the criminal defendant’s most important goals—minimizing punishment once guilt has been ascertained.

One reason that conflicts of interest persist in indigent defense is also obvious; if attorneys for the indigent are to be paid at all, they must be paid by someone other than their clients. The resulting division of loyalties is clearly undesirable, but what is the alternative? In this Article we suggest that realistic alternatives exist and explore ways in which they might be implemented.

Our proposals for restructuring indigent defense are not intended simply as a theoretical response to a theoretical problem. Criminal defense systems seem to be in a state of perpetual crisis. The grave inadequacy of existing systems for serving the indigent is widely acknowledged and widely discussed. We offer far-reaching suggestions for change, not just because they are analytically interesting, but because the results of
existing indigent defense methods are often abysmal and because the need for effective reform is acute.

In focusing on conflicts of interest as a critical source of present difficulties, our diagnosis and our proposed solutions differ in three fundamental respects from those which are common to the existing literature on reform of indigent defense. First, though we are aware of the importance of resource levels, and will comment on their relationship to our analysis, our approach largely takes as given the resources allocated by prior political decision to indigent defense. We seek to show that at any level of resources, reorganization of an indigent defense system can produce gains for both the criminal defendant and society as a whole. Second, we place little store in reliance on case-by-case litigation of ineffective assistance claims or in efforts to strengthen doctrinal formulations of the Sixth Amendment standard. Though doctrinal change could probably improve the quality of indigent defense services to some extent, claims of ineffective assistance on the record of a particular case can have little influence on the overall operations of an indigent defense system. And many areas of indigent defense practice will remain impervious to any conceivable modifications of the constitutional test for effective assistance.

A third difference is the most basic. Contrary to the prevalent approach in the literature, we do not take as our paradigm a large defender organization providing the lion’s share of indigent defense services for a city or county, and we do not focus on efforts (desirable though they may be) to write charters that attempt to guarantee such organizations legal independence from the government that funds them. Efforts to promote formal, legal independence for defenders have recently become the subject of acute controversy in the federal courts. An ad hoc committee, appointed by the Chief Justice to review federal defenders services, has recommended creation of a Center for Defender Services independent of the federal judiciary. However, the Judicial Conference of the United States and most federal defenders oppose that concept on the ground that it would leave defender budgets even more vulnerable than when, as now, they remain under the administrative aegis of the Judicial Branch. We differ from both sides in this debate because we see budgetary vulnerability and implicit conflicts of interest as inherent in the large defender model, whatever its legal and administrative status. Our alternative paradigm is that of a free market for defense services, one that would, so far as possible, function in the same way that the existing market functions for affluent defendants able to retain their own counsel. Rather than stressing efforts to strengthen the formal independence—and monopoly position—of established Public Defender agencies, we seek, in a word, to “privatize” the delivery of indigent defense services.

Compared to other public functions such as education or the military, indigent defense remains a tiny sector of the economy. Yet few would consider it unimportant. The significance of effective criminal defense for safeguarding basic liberties and protecting the innocent needs no elaboration. We will not speculate here about why conservatives and libertarians who have actively promoted deregulation and privatization as means to restrain government power and enhance consumer welfare in other sectors of the economy have not heretofore sought similar reforms for the indigent defendant. Nor will we suggest that those who oppose privatization in other areas, such as education,
logically must adopt the same position with respect to indigent defense. The difficulties of privatization in each area are distinct, and potential gains are not always comparable. We propose to address on their own terms the organizational problems of indigent defense and to suggest a fresh approach to understanding and correcting them.

Part I of the Article analyzes the structure of the attorney-client relationship and identifies the problems that contractual or institutional arrangements must seek to minimize. Part II describes existing methods for the delivery of indigent defense services and assesses their ability to address these problems. Part III develops alternatives to existing arrangements. We consider three groups of approaches: insurance models, deregulation models and voucher models. We conclude that insurance models, though theoretically revealing, are impractical. Deregulation models offer an immediate, easily implemented but partial solution. The more ambitious voucher models, adapted to local conditions in various jurisdictions, provide a practical and effective cure for many of the major ills of indigent defense organization, to the ultimate benefit of both defendants and the public at large.

I. GOALS AND PROBLEMS IN THE ATTORNEY-CLIENT RELATIONSHIP

Criminal defendants, we may assume, are ordinarily interested in winning acquittal or, if that fails, the lowest possible sentence. If they must support their own defenses, they will also prefer to achieve these goals at the lowest possible cost. Defendants facing substantial prison terms will spend large sums to produce even small increases in the chance of acquittal, but at some point diminishing returns presumably prompt most defendants to economize on the expenditure of their own or their family’s resources. Conversely, defendants of moderate means may run out of funds while a potentially productive defense effort remains unfinished; they may regret the inadequacy of their available savings.

Criminal lawyers, whether assigned to indigent defendants or retained by affluent ones, must make hard choices—including decisions about how much work to do (whether to investigate factual leads, research legal issues and file particular motions) and about what advice to render in matters of judgment (whether to recommend accepting a proposed settlement, holding out for a better offer, or going to trial in hopes of an acquittal). For all of these decisions, the lawyer’s personal interest may diverge from that of her client. Thus, under the best of circumstances, the relationship between defendant and counsel involves serious agency problems. In the case of retained counsel, these problems are mitigated by the fact that the lawyer must attract and keep clients, and will do so by creating and maintaining a reputation for serving the clients’ interests even when they conflict with his own. The indigent defendant has no such protection. His counsel is chosen not by him but by the judge, the public defender’s office, or some private organization which contracts with the government to provide attorneys for the indigent. If the attorney wishes future cases, she must indeed maintain her reputation, but only with those who provide her with business, not with potential defendants themselves.

The attorney-client relationship thus poses three sorts of problems—those involving incentives for the attorney to act in her client’s interest (incentive problems), the need for information about the quality and loyalty of alternative providers of defense services
(information problems), and protection against the risk of unanticipated need for criminal defense services (insurance problems).

A. Incentive Problems

Incentive problems are two-sided. If the lawyer’s fee is based on an hourly rate set at a figure that is low, relative to the lawyer’s other opportunities, or if total resources available for the case are too meager, attorneys may forego useful investigations and may avoid trial even when there are good chances for acquittal. If hourly fees are too generous and if available resources are unlimited, attorneys may pursue unproductive investigations or hold out hopes for acquittal at trial when a guilty plea would better serve the client’s interest.

The first problem is not confined exclusively to the indigent. Middle-class or even affluent clients often lack resources sufficient to assure their attorneys a compensatory fee if a fully litigated trial turns out to be the defendant’s best option. Conversely, the second problem is not experienced exclusively by the wealthy. Attorneys for the indigent also may do unnecessary work and run up unjustifiably high fees if their opportunity costs are low and if state support is not effectively restricted. Here, optimal behavior is distorted by a conflict of interest between the defendant and his lawyer on the one hand, and the taxpayers who pay the bill on the other. As in any situation in which the choices of a buyer and seller are supported by a third-party payor with imperfect monitoring capabilities, expenditure is likely to skyrocket. Health care is the classic case in point.

A further complexity, in the case of indigent defense, is that high fees and imperfect third-party monitoring generate simultaneous conflicts in three directions. When the attorney must decide whether to pursue a marginal or worthless investigative effort, the client is largely unaffected, and the conflict of interest is between the attorney and the state. But when the decision concerns whether to recommend rejection of a plea offer, the attorney’s personal interest in gaining lucrative fees by going to trial may clash with those of his client, who wants to avoid a harsh post-trial sentence. At the same time, the client’s interests in minimizing the expected sentence also diverge from those of the state, which may want to minimize the expense of the representation, subject only to the constraints of not deliberately convicting the innocent or affording constitutionally ineffective assistance.

Contractual and institutional arrangements are necessary to deal with these problems of distorted incentives and conflicting interests.

B. Information Problems

In order for anyone—judge, state government, or defendant—to choose the best provider of defense services, he must have information on what will be provided. This is a particularly serious problem for the defendant, since he may have had little previous experience with the criminal justice system. The poor may be especially disadvantaged in this regard, since they generally have less access to lawyers and other sources of information about professional competence. Yet the indigent defendant is more likely
than a middle class defendant to have faced charges before or to know someone who has. And the professionals with whom a middle class defendant has contact may have little accurate information about the criminal defense bar. Indigent defendants and the more affluent face different but perhaps equally serious barriers to informed choice.

The information problem is less serious if the attorney is chosen by a judge, by a public defender allocating cases to lawyers under him or by a state agency contracting with an independent provider of defense services. In each of these cases the decision is made by someone who has close contact with the criminal justice system and an opportunity to observe the quality of service rendered by alternative providers. Yet the incentive and information problems are in tension. The defendant has the incentive to choose a vigorous, effective advocate but may lack the information to do so. A public official who chooses for the defendant is likely to have better information but a weaker incentive to make the best choice. Indeed, as we have seen, the official, appraising an attorney’s ability from the standpoint of the court system, has incentives to value cooperativeness, disinclination to work long hours, and other qualities that might not win favor with defendants themselves. Providers may end up being selected according to how well they serve the court, not how well they serve defendants.

**C. Insurance Problems**

Potential criminal defendants—which is to say, all of us—face the risk of having to incur the very high cost of an effective criminal defense. Being accused of crime is not wholly dissimilar to catching a potentially incapacitating or fatal disease. Attempts to combat the problem can be enormously expensive and, in the end, may or may not prove successful. A large share of personal and family resources may be consumed in the effort. Not surprisingly, health insurance to spread the financial risks of catastrophic disease is widely available through the market. Yet insurance against the financial risks of becoming a criminal defendant is not. One function of a public defender system is to provide a substitute for the nonexistent insurance. Public funds are available only to the “indigent.” But middle class or even wealthy individuals can easily be rendered indigent by the costs of defending against a serious criminal charge. When the affluent defendant runs out of funds, he can qualify for appointed counsel, either to complete pretrial and trial efforts or to pursue an appeal. The economic effect is comparable to that of an insurance policy with a very high deductible.

In considering how different institutions perform the insurance function, we must distinguish between two sorts of uncertainty:

1. Uncertainty as to whether someone will be arrested and on what charge.

2. Uncertainty as to how complex the case will be.

The second sort of uncertainty requires further explanation. By a complex case, we mean a case in which additional expenditures on defense continue to provide substantial benefits to the defendant up to a high level of expenditure. A simple case is one in which additional expenditures above a fairly low level produce at most small benefits for the
defendant. Simple cases include both those in which the prosecution’s case is so weak that defense expenditures are almost unnecessary, and those in which it is so strong that defense expenditures are almost useless.

We find it useful to distinguish between two sorts of uncertainty regarding the complexity of the case. They are:

2a. Uncertainty that can be resolved before the attorney is chosen.

2b. Uncertainty that can be resolved only after the attorney is chosen.

The various kinds of uncertainty affect the relative advantages and incentive problems of different kinds of pay-outs that an insurance program might afford. Three basic pay-out methods may be distinguished: lump-sum payments, variable (fee-for-service) payments and in-kind payments.

In the lump-sum payment approach, the insurance policy pays a fixed amount or, more commonly, one of several fixed pay-outs, depending on which of several risks materializes. Lump-sum payments are common in disability insurance and certain kinds of supplemental health insurance (coverage for school children, for example). The lump-sum system is also common in indigent defense; as we shall see, many jurisdictions pay appointed counsel a flat fee per case, with different amounts often specified for misdemeanor, felony and capital cases.

Variable (fee-for-service) payouts are probably the most common form of health-insurance coverage, and this system is also used in indigent defense; some jurisdictions compensate appointed counsel on an hourly basis for all reasonable effort both in and out of court. Fee-for-service pay-outs also exist in some commercial policies for reimbursing reasonably necessary counsel fees incurred in defending against civil claims. This form of coverage appears to be uncommon in civil liability coverage, however.

In-kind pay-outs are the predominant form of coverage in health insurance provided by the Veteran’s Administration and Health Maintenance Organizations (HMO’s) and in pre-paid legal service plans available through unions or employers. In commercial insurance against civil liability, the insurer typically undertakes to defend against any covered claim, using its in-house legal staff or selecting outside counsel at its sole expense. In some areas of malpractice or products liability insurance, the value of defense services provided in-kind is often greater than any indemnification for damages. The in-kind payment system is also the dominant form of criminal defense “insurance” in jurisdictions that rely on a public defender.

Variable (fee-for-service) pay-outs afford the most complete coverage for the insured, but present large incentive problems. The insured and the service provider have only weak inducements to control costs, and monitoring by the insurer may not be fully effective, as escalating health-care costs have made clear. Lump-sum payments avoid the monitoring problem for the insurer (at the cost of possible overpayment on some claims) but leave the beneficiary self-insured for the risk that providers will be unwilling to take on his case because they readily identify it as an exceptionally complex one that cannot be
treated for the lump-sum fee. The problem is different when a complex case cannot be identified as such before a service provider accepts it. In that instance the lump-sum may be adequate to induce a doctor or lawyer to commit to providing the necessary services. The risk of unforeseen complexity then shifts to the service provider, but because the lump-sum fee affects his incentives, the monitoring problem is transformed. Surveillance, directly or through reputation, is no longer necessary to prevent excessive provider services but is now required to assure that services are sufficient, and the responsibility for monitoring shifts from the insurer to the insured.

In-kind pay-outs also solve the monitoring problem for the insurer and, unlike lump-sum payments, they protect the insured against the risk of complexity that a service provider could detect at the outset. Their disadvantage is the same as that which the insured faces under a lump-sum payment when complexity is initially disguised. The service provider bears the risk of exceptional complexity, but monitoring by the insured is essential to assure that adequate service is provided.

In-kind payout of defense services in civil liability insurance presents a significant variation. As long as the insurer must cover the full amount of any damage award, the insured escapes the need to monitor the adequacy of defense services. But if insurance coverage is less than the potential damage exposure, the insurer has only limited incentives to render adequate service, especially on litigation effort that serves only to reduce the damage award. In settlement decisions, conversely, the insurer with limited damage exposure may have an incentive to choose “excessive” service. Assume that the plaintiff has a 50% chance of establishing liability, in which case he will recover one million dollars at trial. If policy coverage is limited to $500,000 and the plaintiff offers to settle for that amount, the settlement is likely to be in the joint interest of the insurer and the insured defendant, and is certainly in the private interest of the defendant, who has nothing to gain and $500,000 to lose by going to trial. Yet the insurer, who ordinarily has legal control over the decision to settle, can expect to gain by going to trial, provided that litigation expense will be less than $250,000. Here monitoring by the insured, though essential, may be impeded by the difficulty of extracting sufficient information bearing on the decision whether to risk trial. Perhaps reflecting courts’ concerns with the weakness of reputation and other monitoring mechanisms here, cases increasingly have permitted after-the-fact suits by the insured against the insurer for wrongful refusal to settle.

The different mix of advantages and drawbacks in each pay-out method helps explain why all three approaches are found in most forms of insurance for legal and medical services. Lump sum, variable and in-kind pay-out packages co-exist in the market, and the insured can select the pay-out system that best suits his own situation. In one respect, however, indigent criminal defense is an exception. As we shall see in the next Section, lump-sum, variable and in-kind approaches are all important forms of indigent defense “insurance,” but jurisdictions tend to choose one or the other, or to use them in a specified combination. Nowhere is the insured, the indigent accused, permitted to select the package that best meets his own needs.
II. The Present System

A series of Supreme Court decisions mandates publicly-funded defense for indigent criminal defendants, but not the institutional form of that defense. Existing methods are of three basic types: public defender programs, contract defense programs and assigned counsel programs. In this section we describe these approaches and consider the extent to which they successfully address the problems of incentives, information and insurance.

A. Public Defender Programs

In a public defender program, an organization staffed by full-time or part-time attorneys represents nearly all indigent defendants in the jurisdiction.

In most jurisdictions the defender organization is an agency of the executive branch of state or county government, and in more than half the others, the public defender is an agency of the judiciary. A minority, roughly 10–15% of the defender offices, are organized by private non-profit corporations, which perform the defender function under contract with the city or county.

Contrary to the prevailing popular conception, most defender offices are small. Some 75% employ three or fewer full-time attorneys, and at the time of a 1984 study only sixteen employed more than fifty full-time attorneys. But the public defenders in Los Angeles and in Cook County (Chicago) each employ more than 400 attorneys. Even the small defender offices usually employ investigators, and the larger offices tend to have substantial support staffs, including social workers and paralegals.

Though all defender systems are funded directly or indirectly by the government, there are significant differences in the government’s formal control over the Chief Defender. Usually county officials appoint the Chief Defender, but in some places he is appointed by a bar association committee, by judges (especially when the defender office is an agency of the judicial branch) or, in the case of a Community Defender, by the board of the non-profit corporation. Public Defenders are elected in Florida and in parts of California, Nebraska and Tennessee. Election of the Defender guarantees his independence from county government and the court, but at the cost of accountability to voters who may not regard acquittal or early release of criminal defendants as a top priority. In Florida, some Defenders apparently have won election by pledging to cut the office’s staff and expenses.

The various selection methods do not necessarily preclude appointment of Chief Defenders who will guard the independence and resource needs of their offices. Nearly all Defenders are philosophically committed to protecting the indigent. Some have aggressively challenged defective arrangements, by declining to accept new cases or suing the court system for inadequate financial support. Defender staffs have sometimes gone on strike to protest excessive caseloads, which the defenders felt were forcing them to render inadequate service. Still, most Chief Defenders temper their zeal with pragmatic instincts for bureaucratic survival; if they did not, they could not keep their jobs. Thus, for most Defenders, most of the time, accommodation to the case management and
budgetary priorities of the court and county government is a fact of life. And as a result, the great majority of defender systems are understaffed and underfunded; they cannot provide their clients with even the basic services that a nonindigent defendant would consider necessary for a minimally tolerable defense.

Public defender programs tend to be concentrated in urban jurisdictions. Nationally, only about 33% of all counties have a public defender, but 80% of the fifty largest counties do, and 65% of the population lives in counties with a public defender program. Public officials generally view a defender as the most cost-effective approach when case volume is high. Other advantages claimed for a defender system are that its full-time attorneys can become experts in criminal trial practice, and that their salaried position affords them greater independence from judicial pressure. Critics of the defender approach argue, however, that defender offices tend to become overwhelmed with cases, that a defender system excludes the private bar from involvement with criminal law issues, and that the program “will grow to become a large, bloated bureaucracy.” Critics also object that defenders “file too many motions and clog the docket because they are on salary and do not have to account for their time.” This last objection simply restates in negative terms the very qualities that proponents of the defender consider its advantage; what some see as independence is for others irresponsibility.

As a solution to the problems of incentives, information and insurance considered above, the defender approach is plausible but imperfect. The information effects are straightforward. Subject to his budget constraints, the Chief Defender can hire the best attorneys possible, and can know their abilities first hand before assigning them to cases. He is probably more able than the defendant to select the best attorney for the case, at least if the meaning of “best” is unambiguous. But if the Chief Defender values attorneys for their ability to move cases quickly and to persuade reluctant defendants to plead guilty, the accused might be better off making his own, poorly informed, choice. This problem is not lost on the supposedly unsophisticated defendants whom the public defender ostensibly protects from exploitation in the market. Indigents commonly mistrust the public defender assigned to them and view him as part of the same court bureaucracy that is “processing” and convicting them. The lack of trust is a major obstacle to establishing an effective attorney-client relationship. The problem was captured in a sad exchange between a social science researcher and a prisoner: “Did you have a lawyer when you went to court?” “No. I had a public defender.”

The twin incentive problems are to assure that defenders do not slight the client’s interest in adequate service or the taxpayer’s interest in controlling costs. The latter concern is met directly by government power to fix the defender budget and its control or influence over the choice of the Chief Defender. The Chief Defender, in turn, may lobby for more resources (just as the District Attorney might), but once the appropriation is determined, he will be forced to insist that his staff allocate time and resources carefully to provide the best possible service to the clientele as a whole, within the limits of budget constraints.

The other incentive concern (not slighting the client’s interest) is more problematic. One might ask why the Defender or his staff would bother to do anything for their clients,
beyond the minimal effort required to avoid professional discipline or suit for ineffective assistance. One answer is personal pride and a commitment to professional values. Many defender offices develop an esprit de corps, in which they view acquittals as victories and severe sentences as defeats in a continuing competition with the prosecutor’s office. The agency might be staffed and run by idealistic lawyers, trying to do their best for the poor and unfortunate.

To the extent that idealistic motivations are operative, the defender approach provides a distinctive way to reconcile the twin incentive problems. When government controls compensation case-by-case, as in the assigned counsel systems considered below, its need to prevent excessive service is at every step in direct tension with the defendant’s need to assure adequate service. In the defender approach, the state exercises its cost control function wholesale, leaving the monitoring function at the “retail” level to the Chief Defender and other supervisors in his office. Their annual budget leaves them (like prosecutors) the flexibility to invest enormous resources in a particular case if their sense of justice requires, free of the chilling effect of case-by-case external review. But even when mediated in this way, the cost control function constrains the management of nearly all cases nearly all of the time. The annual bottom line may even create a more powerful and pervasive cost-control ethos than would exist for a private attorney who had to justify a single claim for fees in an individual case.

Considerations of narrower self-interest may join with idealism in providing incentives for adequate service. To win the esteem of colleagues, adversaries and judges, and to pave the way for subsequent career moves, the staff attorney needs a reputation for vigor and effectiveness. The reputation effect can operate powerfully at trial but is unlikely to constrain an attorney’s low-visibility decision to recommend a time-saving plea. The reputation effect may even distort his advice by inducing him to recommend trial in a case that would be a “good vehicle” or to plead out some defendants in order to permit better preparation in high-visibility cases. In any event, self-interested reasons for effective performance, as reinforced by idealism and office esprit de corps, must compete with office attitudes that run in the opposite direction—that of restraining costs and cooperating in the court’s desire to move cases. The adversarial attorney thus may lose collegial esteem or the Chief Defender’s approval as a result of vigorous efforts. In one recently publicized case, the Atlanta Public Defender demoted a staff attorney because she had filed a motion asking the local judges to appoint her to no more than six cases per day.

The insurance problems are a function of the incentive issues just canvassed. Like any insurer which provides an in-kind pay-out, the defender has in-house control to prevent excessive effort, but it bears the risk of unforeseen complexity, and the insured (the accused) must monitor performance to prevent short-cuts and inadequate service. In one respect the criminal defendant is better placed to control counsel’s effort because the decision whether to settle is legally his alone to make; the insured defendant in civil litigation often has no such protection. On the other hand, the criminal defendant has less capacity to assess litigation risks than many civil defendants, usually hospitals or manufacturers with their own legal staffs.
An alternative possibility for monitoring is the after-the-fact suit for malpractice or constitutionally ineffective assistance, roughly analogous to the civil defendant’s suit for an insurer’s wrongful refusal to settle. But the malpractice suit is virtually a non-existent remedy for the criminal defendant. An ineffective assistance claim is almost equally improbable as a monitoring device. First, many departures from fully adequate service do not rise to the level of constitutionally ineffective assistance. The constitutional standard is low, and what the defendant wants to assure is not just a minimally adequate effort but the effort that an attorney with the right incentives would provide. In addition, the severe penalties that can follow conviction at trial mean that an attorney’s recommendation to plead guilty can almost never be proved unreasonable, however much it may be influenced, consciously or subconsciously, by resource constraints. Finally, ineffective assistance claims often can be brought only in post-conviction proceedings and such claims must be brought in a post-conviction proceeding when conviction is on a guilty plea; thus the defendant’s only tool for monitoring is one he must invoke without a constitutional right to professional help.

The weakness of available incentives to assure adequate services and the absence of effective after-the-fact monitoring leave the Public Defender as a highly flawed solution to the incentive, information and insurance problems. Although idealism undoubtedly motivates many defenders to seek the best outcome for their clients, the system as a whole is driven by political goals which often conflict with that objective. A court system troubled by full dockets and high crime rates may well decide that lawyers with an idealistic commitment to getting their clients acquitted, a strong aversion to guilty pleas, or a determination to assure the lowest possible sentences are not the lawyers it wishes to put in charge of indigent defense.

B. Contract Defense Programs

In a contract defense program, individual attorneys, bar associations or private law firms agree to handle a specified volume of indigent defense cases for a specified fee. Although a contract defender could in theory devote all his time to indigent defense work, contract defenders almost invariably maintain a substantial private practice. Unlike the public defender, a contract defender normally handles only a part of the jurisdiction’s indigent defense caseload, and counties that use this approach may have several independent attorneys or firms under contract. Contract defender programs are becoming more popular, but nationally only about 10% of all counties use this type of program as their primary system for delivering indigent defense services. Many others, however, use the contract method as their back-up system for cases that the public defender cannot accept.

Two types of contracts are common. In the global fee approach, the contract defender agrees to accept all cases of a certain type—for example, all felonies or all juvenile cases—for a single annual retainer. Many county officials prefer this approach because it keeps the indigent defense budget predictable and puts a cap on total expenses. That leaves the contract defender with the risk of unforeseen increases in caseload. In effect he is selling the county not only legal services but insurance. The contract defender has much less ability to control the court’s caseload, which is largely a function of the
The other common approach is the individual fee contract, in which the attorney or firm agrees to handle a specified number of cases for a flat fee per case. These contracts protect the attorney from the risk of unforeseen increases in case volume, but he still bears nearly all the risk of unforeseen complexity. More than half the contract defender agreements use the individual fee approach.

The contract method is attractive for jurisdictions that are too small to support a full-time public defender. It has become the predominant system in some sparsely populated areas, such as North Dakota. Costs tend to be lower than in assigned counsel programs, and the quality of service may be superior; as compared to assigned counsel, contract counsel have more opportunity to specialize in criminal practice. Critics of contract programs fear that counties will award contracts to the lowest bidder, without regard to attorney qualifications or the proposed level of service. Apparently this has not happened in any blatant way to date; competitive bidding is widely used but in nearly all cases surveyed, the cost per case was set in advance, and bidders were selected on the basis of quality. Critics also charge that global fee and individual fee contracts have induced attorneys to stint on their time and on the back up services of investigators and experts.

Information, incentive, and insurance problems arising in contract defense programs largely parallel those in the public defender service, but with several important variations. The indigent is insured against exceptional complexity but, as in a public defender program, the accused bears the burden of monitoring. Effective tools for carrying out this function are largely absent.

The information problem in a contract system presents itself in two stages: officials must award contracts to attorneys and then assign individual cases to one of the previously designated contract recipients. Often the first decision is made by county government and the second by a court administrator or by the judge responsible for the case. At both stages, officials are in a good position to evaluate attorney competence. Indeed, competitive bidding focused on quality of service offers a powerful vehicle for ascertaining what qualifications and support services are available through competing providers. And compared to some assigned counsel programs discussed below, there is more prospect that officials will use their superior knowledge in the right way, because the county’s defense costs are not affected by the choices made either at the case-assignment stage or at the initial contracting stage—at least when the contract price is fixed in advance and excluded from negotiation or competitive bidding. For whatever it chooses to spend, the county should be in a position to choose the best available defenders. There is one qualification, however. Although defense costs are independent of which attorneys are selected, total court costs are not. Thus, officials might hesitate to choose attorneys known for filing many motions, driving hard bargains or insisting on trials, even if the lawyers offered to provide these services at no extra charge.

Contract programs, like public defender programs, address only one side of the incentive problem. Because fees are fixed, either per case or per annum, attorneys have a powerful
incentive to avoid *unnecessary* service, but there are few direct incentives for *adequate* service. Indeed, global fee and individual fee contracts give the attorney a powerful disincentive to invest time and resources in his indigent cases. In this respect the contract attorney poses a much larger problem than a public defender does. Public defenders have few direct economic incentives for adequate service, but apart from any taste for pure leisure, they have no concrete *dis*incentive to work on their cases. Public defenders may cut costs on some cases to free up resources for others, but they cannot take home unspent cash at the end of the year. The contract defender, in contrast, is in business for a profit. Money saved on defending one case need not be spent on another; it may simply enlarge the Christmas bonus. Perhaps worse, time saved in handling indigent cases is freed up for more lucrative business, and an overly busy attorney is unlikely to turn away paying clients when he has the alternative of cutting low-visibility corners in his indigent case commitments. These dangers are intrinsic to all contract defender programs and have produced seriously deficient service in many.

A countervailing incentive for adequate service in contract programs is professional pride and ideological commitment. But we would speculate that the latter factor, though important for public defenders, operates less strongly among contract attorneys, who may see their arrangements primarily as ways of attracting business to a profit-oriented practice. Another factor that may encourage adequate service is the opportunity to renegotiate contracts annually. Proponents of the contract system argue that the continual prospect of service providers leaving the market puts pressure on counties to keep contract prices at a compensatory level; an institutionalized public defender has less capacity to resist steady increases in its caseload staff ratio. A public defender’s leverage would be powerful if its staff chose to strike, but defender strikes have proved problematic, legally and politically.

A final inducement to providing adequate service may be the attorney’s desire to win renewal of the contract. But contract renewal requires pleasing the county, not the client, and the county’s priorities are, as always, ambivalent. Over the long run, the lawyer may attract more business and generate more profit per case by keeping costs, and hence next year’s fees, lower than those of potential competitors. Cost-cutting is not undesirable per se, but it is problematic in indigent defense programs because the client who suffers from reductions in the quality of service has no control over the flow of business to the attorney who benefits from the decision to cut costs.

Though the contract system resolves some incentive, information and insurance problems better than a public defender does, overall its structure is seriously flawed. The existence of competing service providers in the contract system should be advantageous, but the potential benefits are lost because court officials rather than clients control the flow of cases to the attorneys.

*C. Assigned Counsel Programs*

In an assigned-counsel program, a member of the private bar is appointed on a case-by-case basis for each criminal defendant. About 60% of American counties use assigned counsel as their primary method of assuring indigent defense, and nearly all the others
rely on assigned counsel for cases in which public and contract defenders are disqualified or unavailable.

The judge responsible for the case or another court official usually makes the assignment decision. Sometimes the selection system is entirely informal, and appointments are distributed ad hoc to attorneys the judge knows or to those who happen to be present in court. More typically, the assigned attorney is chosen from a list established in advance by either the court, the local bar association or each judge for his own cases. The judge may use a formal, predetermined rotation plan to select an attorney for each case, or he may make a selection from the list less systematically. All members of the bar may be eligible for the list, or there may be a few simple prerequisites, such as a certain number of years of experience. About a quarter of the assigned counsel jurisdictions have more elaborate eligibility requirements, a system to screen applicants for inclusion on the list, provisions for monitoring and ancillary support services.

Desire to serve is not necessarily a condition of eligibility. Nearly all courts have authority to appoint an unwilling attorney, and such a power is probably an essential backup for cases that involve extensive conflicts of interest or an extraordinarily unpopular defendant. But in many jurisdictions, conscription of unwilling attorneys is a routine feature of the assignment system: all eligible attorneys are included on the list, and they are obligated to serve when called.

A variety of compensation systems are used in assigned counsel programs. In some, attorneys receive a flat fee per case or per appearance, usually with different amounts specified for juvenile cases, misdemeanors and felonies. Other jurisdictions pay on an hourly basis, often with one rate for time spent in court and a somewhat lower rate for time spent in preparation.

Hourly rates vary from low in some jurisdictions to derisory in others. A March 1992 survey found many jurisdictions still paying only $10 or $20 per hour, rates that are inadequate even to meet the attorney’s office overhead. Low rates are not exclusive to Southern or mainly rural states. Hourly rates for out-of-court time stand at $25 for New York, $20 for Connecticut and $15 for New Jersey.

When compensation is on an hourly basis, judges or court administrators always review fee claims, and they often reduce fees, sometimes in a way that the attorneys consider arbitrary. In any event, there is usually a cap on total compensation payable. In the federal courts, the cap, now $3500 for felonies, may be waived, though only by the chief judge of the circuit court of appeals. In many states the cap may not be waived, and as caps are low, compensation systems that are ostensibly based on an hourly rate become flat-fee systems in practice.

The low caps reported in several surveys in the 1980’s have been raised to some extent. But as of March 1992, the maximum fee for a non-capital felony was still only $1250 in Illinois, $1200 in New York, $500 in Oklahoma, South Carolina and New Hampshire, $400 in New Mexico and $350 in Arkansas. In Kansas, one of the few states to set different caps for plea and trial cases, $1000 is allowed for felonies that go to trial; the
maximum for a guilty plea case is $400 for the most serious cases and $250 for less serious felonies. In Virginia, compensation may not exceed $265 for felonies carrying a potential sentence up to twenty years; the maximum for felonies punishable by sentences over twenty years is a mere $575—enough to fund less than three days’ work at the modest authorized rate of thirty dollars per hour. Finally, some jurisdictions regard indigent defense as a “pro bono” obligation, and appointed counsel, usually conscripts rather than volunteers, receive no compensation at all. Although the no-compensation approach is exceptional, flat fees or fee caps are so low in many jurisdictions that hourly compensation in cases that go to trial is virtually nil.

Assigned counsel programs are often defended on the ground that they provide good training for new attorneys, that private counsel are more independent than public defenders and that private counsel develop better rapport with their clients, who often mistrust the public defender. Such programs can also be less expensive than a defender system, at least when fee caps are low and attorneys are expressly or indirectly conscripted to participate. Others see the same points as disadvantages. Low rates draw mainly inexperienced or less effective lawyers and discourage preparation; yet when compensation is anywhere near market rates, costs become unpredictable and much higher than those of a public defender.

In terms of the information, incentive and insurance problems we have canvassed, assigned counsel programs pose numerous obvious problems. Judges and court officials who select counsel have the ability to acquire good information about attorney effectiveness, but they have little incentive to acquire such information and even less reason to act upon it. Their own interests are best served by assigning an attorney known to be cooperative rather than aggressively adversarial.

With respect to the attorney himself, the incentive problem is to provide inducements for sufficient but not excessive effort. Low hourly rates, low fee caps and mandatory service pro bono nicely solve the latter half of the problem but leave the assigned attorney with powerful reasons to minimize the time and effort devoted to the case. The more generous states, a small minority, face different problems. Hourly rates close to market levels and an absence of fee caps give the right incentives for adequate service but they risk unnecessary attorney effort and excessive cost. Most of these more generous jurisdictions rely on reputation effects, along with case-by-case review of attorney fee submissions, to provide the right cost-control incentives, but monitoring of this sort is expensive and not always successful. And even if not abused, a program of compensation at near-market rates puts unpredictable budget demands on the county and tends to cost more than specialized contract defenders or a public defender system.

In terms of insurance problems, the compensation structure is crucial. If fees are paid at near-market levels, the county is in effect self-insured for both the risk of unusual case complexity and the risk of unforeseen increases in case volume. Except in the smallest rural counties, the former risk is spread over a universe of cases large enough to make self-insurance viable. The defendant escapes most of the need to monitor the adequacy of service, at least if he can assume that the assigned attorney has no hidden motivation to cut costs. But the county has an intense need for monitoring to prevent excessive costs.
And since the county may meet that need by assigning attorneys predisposed to be cooperative, or not reappointing those who seem too diligent, the defendant still needs—but largely lacks—some vehicle for effective monitoring of the adequacy of service. In fixed-fee and low-cap systems, the county still bears the risk of unexpected increases in case volume, but the assigned attorney now bears the risk of unusual case complexity, and the burden of monitoring now falls entirely on the party least able effectively to protect his interests—the indigent accused.

III. ALTERNATIVE APPROACHES

In the present section we consider alternative approaches for solving the organizational problems of indigent defense. As we have seen, existing systems resolve, with varying degrees of success, the incentive, information and insurance problems presented for the state, but the indigent’s problems in these areas are left largely unaddressed. There are few reliable mechanisms to insure that attorneys for the indigent vigorously protect the interests of their clients when these clash with the interests of the court system or the government that pays their fees. Before describing specific institutional alternatives, we can help focus the issues by describing three general tools for solving the client loyalty problem, which is the central difficulty each approach must address.

One such tool is to rely on incentives other than individual or institutional self-interest, in particular the attorney’s personal pride, professional ethics and idealistic commitment to helping the poor. This is the solution implicit in all existing institutions. Its power is not negligible, but for reasons already discussed, we believe it is by itself an inadequate counterweight to strong organizational and financial pressures that push in other directions.

While one wants to be sure that institutional reforms do not impair the valuable role of personal and professional ideals, there is a need to supplement idealism with concrete inducements and to diminish the power of countervailing pressures.

A second solution is to use direct incentives to align the interests of defense counsel more closely with those of the defendants. This could be done, within a system in which the state appoints defense counsel, by making reimbursement in part conditional on the outcome of the case, with outcomes more favorable to the defense resulting in more compensation for counsel. But there are at least two problems with this solution—the incentives and the knowledge of those running the program. We want direct incentives because we suspect that the government’s interest is in conflict with the interest of the defendant. Yet there is no simple, objective way to convert successful outcomes into automatically quantifiable bonuses. Setting up a system of discretionary rewards controlled by the state would have a certain air of hiring the fox to guard the chicken coop. If such a reform were introduced, those in charge of administering it could defeat its purpose by paying bonuses to the most cooperative lawyers rather than to the most effective ones.

The second problem is that, even if the system is designed and run with the intention of serving poor defendants as well as possible, those in charge may not have the information
necessary to do so. This is a common problem in institutions designed to substitute administrative rules for market incentives. How a defendant would wish his counsel to trade off the costs and benefits of different strategies is a complicated issue, especially in deciding whether to accept a particular plea offer. Any administrative rule setting the reward as a function of the outcome will represent only a crude approximation of the correct incentives.

A third solution is to transfer the power to select the attorney from the court system to the defendant, introducing consumer sovereignty into the institution of indigent defense. So far as his own interests are concerned, the defendant has precisely the correct incentives. If available information is good enough to allow a defendant to appraise alternative providers of defense services, such a system solves the client’s problem. Even if the defendant cannot judge perfectly among alternative counsel, at least the decision will be made by someone with an interest in making it correctly. Consumer sovereignty is, despite imperfect information, the mechanism that most of us use most of the time to control the quality of the goods and services we buy.

In this section we explore ways to correct systemic malfunctions and enhance the welfare of the indigent by introducing elements of consumer sovereignty into indigent defense programs. Our objective is, so far as possible, to create a system in which defense attorneys do not find it in their interest to sacrifice the welfare of either defendants or taxpayers to their own, nor to sacrifice the interests of defendants to those of the state. Defense attorneys are not the only ones whose incentives will be affected by different institutions for providing criminal defense services. As we will see below, some of the institutions we consider will also affect the incentives of the prosecution. We consider three approaches: insurance models, deregulation models and voucher models.

A. Insurance Models

Probably the ideal way to maximize choice and minimize the risk and incentive problems for all concerned would be to establish a system of private insurance through commercial carriers. The practical obstacles to implementing such a system are formidable; we will consider them in a moment. But it will help clarify the issues if we begin by imagining how a system of private insurance might function.

As we have seen, financial protection against the costs of medical and legal misfortune can be provided in the form of lump-sum, variable or in-kind pay-outs. Companies providing insurance against the cost of criminal defense could offer different policies with a choice among these forms of protection. The customer, in choosing a company and a policy, will wish to insure himself optimally against the expenses of criminal defense. Of course, commercial insurance, even if available, would not entirely solve the problem for the indigent; many of those too poor to finance a defense once accused would also be too poor to buy insurance. The first step, therefore, would be for the state to provide all indigents a voucher to be used for the purchase of insurance.

Each indigent citizen would then select the policy that best met his own needs. Suppose, for simplicity, that he is considering only contracts that offer a fixed sum to pay for
defense against each possible charge. He will want a schedule of payoffs such that the benefit to him of an additional dollar is the same in each situation. As long as that is not the case, the contract can be improved by transferring benefit payments across outcomes, from one where the marginal benefit is low to one where it is high. So the consumer will want a contract that pays large amounts for charges where it is likely that additional defense expenditures will continue to produce large reductions in the probability of conviction up through high levels of expenditure ("complex" cases), and pays small amounts for charges where it is likely that the payoff from defense expenditure drops off at a low level of expenditure ("simple" cases).

The same argument applies to contracts providing a more complicated structure of payments. Insofar as it is possible to distinguish between complex and simple cases involving the same charge, the customer will prefer a contract that provides more money for the complex cases, up to the point at which the marginal payoff is the same in all situations. But sorting of this kind is difficult, and with indigent customers, the insurer could not rely on the common solution of coinsurance, in which the insurer would agree to bear, for example, 80% of the lawyer’s fee and the insured would decide how many hours of legal services he wants. For both variable and in-kind pay-outs, the most likely device for regulating the level of service would be reputation. As in policies of insurance for the costs of defending against civil claims, the customer would decide whether he prefers a firm that guarantees a specific level of service or one that makes no enforceable guarantees but has a reputation of keeping costs down on simple cases and spending whatever is necessary for complex ones.

The argument applies to incentive problems as well. The customer will not only weigh the reputation of the firm but will consider how the payment scheme of a particular contract affects the incentives of the lawyer who will work on his case. The better the contract aligns the interest of lawyer and client, the more attractive the contract.

The insurance approach also would protect the state against the risks of excessive service, unexpected case complexity and unexpected increases in the volume of cases. Each year the state would determine its budget for indigent defense, divide it by the number of indigents in the jurisdiction, and distribute the resulting amount to each indigent as a voucher for purchasing insurance. Insurance providers would control for excessive service and spread the other risks—unforeseen changes in volume and complexity—through re-insurance or their own multi-state operations.

Unfortunately, private insurance against the costs of criminal defense is probably not a practical option. One problem is adverse selection. Some people—mostly those who intend to commit criminal acts—know that they have a higher than average probability of being arrested. To them, an insurance policy covering the legal costs of their defense is very valuable. Other people know that they have a low probability of being arrested. They might still want insurance, but only at a lower price—one reflecting their low chance of collecting. As long as the customers know which group they are in and the insurance companies do not, adverse selection will tend to drive the low risk customers out of the market. The result will be that innocent defendants—the group who, on both moral and economic grounds, we would most like to protect—will be uninsured.
A further problem with criminal defense insurance is informational. Most of us face only a small risk of being charged with a crime. As a result, even if such insurance were available, it would be worth bearing only very small costs in evaluating the contracts offered by competing firms, in order to decide which provided the best package of benefits. Since good information about the quality of such a complicated product is expensive, firms would face only weak pressure to construct well-designed programs.

These theoretical arguments fit the observed facts. Insurance against the legal costs of defending civil cases is a standard feature of liability insurance policies, and of what are usually referred to as prepaid legal service plans. Insurance against the legal costs of serious criminal accusations (other than motor vehicle prosecutions) seems to be nonexistent. Since a viable market for such insurance has not emerged even for non-indigents, there is no strong reason to suppose that providing voucher payments to all indigents would create such a market to serve them. The problem then is to determine to what extent alternative arrangements can provide the desirable features of an “ideal” insurance system.

B. Deregulation Models

A less complicated way to address the organizational problems of indigent defense would be to leave existing delivery systems, including compensation levels and methods of payment, in place but reduce government control over selection of the attorneys who serve. At present, the assignment systems are highly regulated. Public officials normally designate the Public Defender and select contract defenders; a judge or court official determines the necessary qualifications for appointed counsel and screens the bar to select applicants or place conscripts on the list. A public official also assigns one of the eligible attorneys to each case, unless this is done by rotation or at random. The large public role serves both informational and incentive needs. As we have seen, public control can help insure that attorneys are capable and that they do not run up excessive costs. But both effects have mixed consequences. Cost control may occur at the expense of the defendant’s interest in vigorous representation. Quality control may favor cooperative attorneys rather than aggressive ones.

A straightforward solution to these problems is to deregulate the system by reducing government control over attorney eligibility and over the assignment of counsel to particular cases. In a deregulation program of modest ambitions, the government would continue to designate the Public Defender and contract defenders. But screening and eligibility standards for appointed counsel would be eased, and each defendant would select his own attorney from among those eligible. In this section we will first describe how a deregulated system could work and then address the arguments that have been advanced against recognition of defendants’ choices.

1. Implementing Deregulation

The first area for deregulation is the matter of determining which attorneys will be permitted to represent the indigent. We approach this topic cautiously because many observers feel, with good reason, that eligibility standards are already too lax, especially
for serious matters such as capital cases. Yet the nonindigent defendant has a constitutionally protected right to select any attorney who is admitted to the local bar and free of conflict of interest.

Information problems in the selection of competent defense attorneys are undoubtedly serious. These problems may warrant requiring defense attorneys to be certified for criminal practice as a specialty or to meet stiff standards for serving in the most serious cases. Information problems might also justify steps to make details about the qualifications and experience of criminal defense practitioners more readily available. We will not discuss such reforms here, except to note that any justified restrictions on eligibility and any worthwhile efforts to disseminate information should affect rich and poor defendants alike. Court officials might continue to screen attorneys for an indigent defense panel, but inclusion on such a roster should mean only that the attorney is a recommended choice, not a required one. Once general eligibility standards for criminal practice are set, there is no reason for legal regulation to confine the indigent to a smaller pool of providers.

The second area in which government regulation should be relaxed is the matter of selecting the attorney who will represent a particular indigent defendant. So far as we are aware, no American jurisdiction presently permits indigents to select their counsel from among those eligible; many jurisdictions expressly forbid them from doing so. Though such restrictions have occasionally been challenged, virtually every American court considering the issue has held that refusal to accept the indigent’s choice of counsel is permissible and constitutional, even when a defendant’s preferred attorney is a member of the indigent defense roster, has previously been appointed to represent the defendant in other matters and makes clear his willingness to serve.

Efforts to empower the defendant and honor his choices could take several forms. Total deregulation would imply leaving the selection decision entirely to the accused. The attorney selected would have to be willing and able to serve, as in any contractual arrangement, but court officials would play no role in bringing him and the defendant together. At most, the defendant might receive a list of attorneys admitted to the bar or a phone number for the bar association referral service. But approaches with such minimal government involvement seem unlikely to further well-informed choices. In addition, they are almost sure to prove logistically infeasible. Speedy trial concerns of both the state and the accused require that the attorney-client relationship be sealed promptly. Moreover, indigent defendants are often jailed prior to trial. They can be permitted to seek counsel by telephone, their families may help, and their cellmates can provide a large pool of current information. Still, many incarcerated indigents will be unable to pursue a quick and successful search for counsel.

These considerations suggest the need for continued official participation in the choice of counsel. Government involvement need not cloud the incentive structure we want to create, provided that the defendant retains effective control over the ultimate selection decision. In one variation, the defendant could be provided several recommendations, or the name of a single attorney likely to accept appointment. An even more cautious model would continue the practice of having the court appoint counsel, but would advise
defendants that they can choose a substitute if they prefer. The right to choose a substitute not only would apply among appointed private counsel but would extend to replacement of appointed private counsel by the public defender and vice versa.

In any of these approaches, the attorney could refuse an undesired appointment; even the public defender could decline on grounds of overloaded staff or conflict of interest. The defendant would also have to exercise his option promptly to avoid inflicting duplicative work and disruptive continuances. But under no circumstances would court officials be permitted, as they are now, to deny a defendant’s timely request for appointment of a qualified attorney who is willing to serve.

In models involving continued official participation, the “bottom line” might appear unchanged—most defendants would probably end up with the attorney that the court initially appointed or recommended. But some defendants would make a personal choice, and even those that did not would benefit. The mere existence of a right to choose would dissipate some of the distrust that now infects many involuntary attorney-client relationships and would for the first time give the appointed attorney a self-interested reason to value the satisfaction of his client as well as that of the court.

The effect on the incentives of the public defender is more complex. Since defender organizations do not seek a profit, they have no direct economic incentive to try to retain clients; a defender office might even hope that clients would flock to appointed counsel and thus ease its heavy caseload. In the short run, therefore, the defender’s incentives, other than idealism and professional pride, might point toward trying to displease clients. The crucial countervailing factor here is that representation by appointed private counsel entails much higher costs per case than does representation by the public defender. If substantial numbers of clients opted out of defender programs, indigent defense costs would skyrocket. In a choice system, therefore, county officials would have strong pocketbook reasons for insisting that the public defender maintain client satisfaction in order to continue attracting cases. Freedom of choice provides a mechanism that would, again for the first time, give defender organizations an economic incentive to serve the interests of their clients rather than those of the court.

2. Arguments Against Freedom of Choice

Courts and commentators have given several reasons for denying the indigent defendant a say in the selection of his lawyer. In Drumgo v. Superior Court (Cal. 1973), the California Supreme Court made essentially three points: first, that judges who know the abilities of local counsel can choose a more able attorney than the defendant can; second, that granting a choice to defendants in an appointed counsel program would be inequitable because defendants assigned to the public defender cannot decide which lawyer will represent them; and third, that constitutional requirements are met by a satisfactory minimum level of competence, even if the defendant’s preferred lawyer would be better.

United States v. Davis (7th Cir. 1979) held that denial of choice was not merely constitutional but desirable and even, in the court’s view, essential. In that case, a federal
defendant sought the appointment of an approved panel attorney who was already
representing him in another matter. The trial court’s local criminal justice plan did not
expressly prohibit such an appointment, but the magistrate refused to make it, indicating
that to honor a defendant’s choice was “not the policy of this district.” The defendant
challenged the refusal not as unconstitutional but simply as an abuse of discretion. The
U.S. Court of Appeals for the Seventh Circuit rejected the argument. Relying on the 1968
version of the ABA Standards on Criminal Justice, the court mentioned several reasons
for considering freedom of choice undesirable as a matter of policy: freedom of choice
would “disrupt . . . the even-handed distribution of assignments”; habitual offenders,
better informed about the abilities of counsel, would have an unfair advantage; the more
experienced lawyers, in heavy demand by repeat offenders, might be unavailable to other
defendants; and choice, if recognized, might lead to undue impositions on particular
lawyers. For these reasons, the court concluded, refusal to honor a defendant’s preference
was not only permissible but “a necessary element of any system to rationally allocate
assignments . . . .” While treating judicial administration concerns as compelling, the
court dismissed the countervailing defense interests as insubstantial: “The defendant has
suggested little more than personality conflicts and disagreements over trial strategy. . . .”

*United States v. Ely* (7th Cir. 1983) reaffirmed *Davis* and enlarged upon its reasoning.
*Ely,* like *Davis,* involved a federal defendant who made a timely request for appointment
of a competent and willing local attorney. The court again held that the refusal to do so
was not only constitutional but desirable. Its opinion upholding the refusal to honor a
defendant’s choice carries particular weight because it was written by a person who is
ordinarily a staunch supporter of contractual freedom, Judge Richard Posner. Judge
Posner began by noting that the compensation paid to appointed counsel is not generous.
He then explained:

> The best criminal lawyers who accept appointments therefore limit the amount of
time they are willing to devote to this relatively unremunerative type of work; some
criminal lawyers, indeed, only reluctantly agree to serve as appointed
counsel, under pressure by district judges to whom they feel a sense of
professional obligation. The services of the criminal defense bar cannot be
auctioned to the highest bidder among the indigent accused—by definition, indigents are not bidders. But these services must be allocated somehow; indigent
defendants cannot be allowed to paralyze the system by all flocking to one
lawyer. The district judge in this case could not, realistically, be required to
arbitrate a dispute between Ely and another indigent criminal defendant who
wanted to be represented by [the same lawyer].

The reasons given for refusing to honor defendants’ choices are in our view insufficient.
The constitutional issue is largely beside the point. Admittedly, the Sixth Amendment
does not guarantee a defendant the best possible lawyer. But the reasons this must be so
are primarily economic and logistical; the best possible lawyer will frequently be too
expensive or too busy to serve. It is far from obvious that the Sixth Amendment should
be deemed satisfied by appointment of a merely competent attorney when a lawyer the
defendant prefers is ready and willing to serve at no additional cost. But even if refusal of
choice is constitutional, why should this policy be considered a necessary or desirable feature of a sensible indigent defense system?

There is little to the claim that the appointing judge can choose more able counsel than can the defendant himself. The importance of rapport and trust for an effective attorney-client relationship is one reason for doubting the proposition. More decisive is the incentive problem. Even if the judge has better information, he has much less reason to make the best possible choice; indeed the judge has an incentive to use his information perversely by preferring cooperative attorneys to vigorous ones. If the claim about judges’ superior knowledge were advanced as a reason for denying choice to defendants of means, we would not give it a moment’s credence. It deserves no more weight with respect to defendants who are indigent.

Several arguments suggest that defendant choice would produce unfairness among attorneys. In the “even-handed distribution” argument, appointment may be seen as a plum that particular attorneys should not be permitted to horde. But remuneration is seldom too generous, and if it is, inequitable distribution is far more likely when judges or other court officials control the largess than when its distribution results from choices by individual defendants pursuing their own self interest. Indeed, defendant choice would tend to force attorneys to compete away excess profits by offering better services to attract clients. The situation most likely to involve excess compensation is that of the inexperienced or ineffective attorney with few other opportunities, and these are just the attorneys that defendants would try to avoid. Defendant choice would tend to skew the distribution of appointments toward attorneys for whom representing the indigent would not be a financial boon.

This leaves the “undue imposition” argument—that good attorneys would be unfairly burdened with the task of serving the indigent—and the related idea, emphasized in Ely, that “indigent defendants cannot be allowed to paralyze the system by all flocking to one lawyer.” But since busy attorneys can refuse additional appointments, no attorney would ever have to carry more cases than he himself considered fair. There is at most the potential inconvenience of an attorney’s having to turn away the request, but successful attorneys always face this problem when their fees are not raised to the fullest or are not known to potential clients. In any event, court officials can easily avoid the logistical problem by advising defendants about which attorneys are currently accepting indigent cases.

Claims that freedom of choice would cause unfairness between defendants seem strained. One argument is that choice should be denied in assigned counsel programs because it is denied to defendants represented by the public defender. But representation by a public defender provides a mix of advantages and disadvantages compared to the services of appointed private counsel; the absence of one feature in a defender system is no argument against making an otherwise feasible improvement in the operation of the assigned counsel program. And a policy denying defendants their choice of counsel within a public defender office is itself open to question. Offices organized to afford each defendant continuous representation by the same defender probably could permit clients a considerable degree of choice in most kinds of cases. Offices committed to the so-called
“zone defense,” which assigns a different defender to each stage of a case, normally would not be able to permit choice of counsel, but such offices would afford other advantages that some clients might consider preferable, such as the assurance that a highly experienced attorney will handle the trial phase of the case. The important point, for purposes of assuring effective choice, is that defendants displeased with the package of services provided by a defender office—including one that did not permit its clients to select a particular defender—should have the option of switching to appointed private counsel, or even to a competing firm of defenders, just as students choose among colleges without expecting to have complete freedom to choose all of their professors.

Other arguments against choice focus on fairness among defendants assigned to private counsel. The best lawyers are a scarce resource, and they must be allocated by some system other than the price mechanism. Ely conceives a defendant’s request as an attempt to deny the preferred lawyer’s services to another defendant who also wants them. Even if this were true, it would mean only a first-come-first-served principle of distribution, not a bad alternative to discretionary government control. But the conflict is largely hypothetical, assuming a unidimensional ranking of lawyer quality and an unambiguous definition of what makes an attorney the “best.” In practice different defendants, will face different tactical problems, and they may therefore prefer attorneys known for different styles and skills. And a crucial determinant of lawyer quality is trust, often based on a prior relationship. Lawyer A is appointed for defendant X, but X prefers lawyer B whom he already knows. Ely forces X to remain with A, on the assumption that a hypothetical defendant Y also wants B’s services. Yet for all we know, Y was previously represented by A and prefers the lawyer that X is trying to get rid of. Ely binds X to A and Y to B, even though each defendant may prefer a swap. As is the case in most other areas of economic activity, allocation by official decision prevents mutually beneficial gains from the exercise of free choice.

A final concern is that in a system of free choice, defendants with poor information would get good lawyers less often than they do now, while well-informed defendants, a group that may include a disproportionate share of repeat offenders, would benefit. But if this is a “problem,” it is also one ubiquitous in economic transactions. Less informed and less careful consumers often do worse than others, but our society does not consider this difference in outcomes inequitable, even for such vital services as health care and criminal defense; hence our commitment to protecting free choice for the nonindigent defendant even though this policy gives the nonindigent recidivist an advantage over the nonindigent first offender. Should we randomly assign medicare patients to doctors in order to be sure that those who are less informed have an equal chance to get the best care? The advantages of choice for maximizing consumer welfare far outstrip any costs to those who do not vigorously protect their interests.

Moreover, less informed defendants are not necessarily worse off in a system that honors the knowledgeable recidivist’s choice. Even when the defendant’s interests are of the sort that the court in Davis sought to trivialize as “little more than personality conflicts and disagreements over trial strategy,” the importance of choice should be apparent. Though a no-choice regime seems to give the inexperienced defendant an equal chance for the “best” lawyer, this is true only over a short run in which the quality of the attorneys
remains fixed. In a system of choice, selection by knowledgeable consumers will tend to attract better attorneys and drive out the ineffective ones. Over the long run, even the poorly informed defendant is therefore likely to do better than he will in a regime that contains no systematic incentive for attorneys to value the satisfaction of their clients, just as all shoppers benefit from the efforts of those who take the time to compare prices and quality across stores. And if Judge Posner is correct, as we believe he is, to note the impact of low fees on the good lawyers, who “limit the amount of time they are willing to devote to this relatively unremunerative type of work,” countervailing incentives to respect the client’s interest become all the more necessary.

Though we find the arguments against choice thoroughly unconvincing, the virtual non-existence of this system anywhere in the United States suggests that powerful reasons of some sort sustain the status quo. One possibility is that logistical impediments to honoring choice are not as trivial as our own analysis suggests. Yet indigents have long been permitted to select their own counsel in England and Wales, and more recently in Ontario, so there seems little basis for belief that such a system cannot be workable. Moreover, the American Bar Association’s opposition to choice, reflected in the 1968 standards on which the Davis court relied, was largely reversed only a year after Davis in the ABA’s 1980 revised standards. More recently, the Judicial Conference of the United States has recommended “an experimental program . . . in volunteer pilot districts in which certain CJA-eligible defendants would get a limited choice of counsel.” We consider this a sensible approach, though the cautious qualifications are probably unnecessary.

The near unanimity of the judicial decisions rejecting freedom of choice might give us pause, were it not for the courts’ obvious stake in preserving the existing rule. The no-choice regime is easy for judges to manage, requires no new administrative procedures, and above all gives courts a direct means of steering appointments to political supporters or to attorneys who do not make excessive demands on the indigent defense budget and on the courts’ time. The steering factor is probably the biggest advantage of judicial control and at the same time its biggest vice, for a judge’s conception of excessive cost may be a defendant’s conception of vigorous effort. An attorney’s tendency to make heavy demands on court time cannot be a proper basis for disqualification, and courts must monitor attorney fee statements in any event. Judicial control over the appointment process may make monitoring easier, but we assume that this advantage would be a manifestly insufficient basis for disqualifying a vigorous attorney who sought to be included on an indigent defense panel; indeed defenders of the no-choice regime have never sought to justify the system on that basis.

We conclude that deregulation, in order to permit freedom of choice in the context of existing public defender and assigned counsel programs, is an easily implemented and important step toward improving incentives and curing malfunctions in the representation of indigent defendants. The next section considers more ambitious reforms that would go beyond free choice for defendants to change the structure of public defender systems and provide greater freedom of organization for the providers of indigent defense services.
C. Voucher Models

The approaches we consider in this section carry deregulation one further step. They not only permit defendants to choose among authorized providers, but they reduce the extent of governmental control over the providers themselves. We intend no sharp conceptual split between the partial deregulation models just considered and the more ambitious voucher models. With the freedom of choice provided by the former systems, clients might begin to request the services of private firms rather than just the lawyers on an indigent defense panel, and lawyers might organize to attract such business. If so, the more extensive changes we now want to consider could quickly evolve from initially cautious forms of partial deregulation. For convenience, we will designate as a “voucher” model any system in which lawyers who serve the poor have freedom to organize their practice as individuals or firms, with or without specialization, and to compete for the business of indigent clients. The voucher would be the guarantee of state payment that the accused can take with him to any individual or group provider of criminal defense services.

Because government would not control the organizational form employed by indigent defense providers, a number of different approaches would be likely to materialize. We would expect them to include solo lawyers, small groups of practitioners and larger firms. Providers would vary not only by size but by kind of practice, just as they currently do in most areas of legal work. Some practitioners might be generalists who occasionally take a criminal case. Most would probably be specialists—in litigation, in criminal practice or even in a particular kind of criminal practice, such as juvenile cases, traffic offenses, or major felonies. These variations already exist among those who represent nonindigent defendants and the increase in the client pool made possible by a voucher system would permit further specialization. We expect that most criminal defense specialists, whether individuals or firms, would serve both poor and affluent clients, though some might specialize in serving the indigent. Finally, we would not exclude the possibility of a government-run staff of salaried public defenders, financed by vouchers collected from clients. A public defender of this sort would not compromise the value of a voucher system, so long as alternate private service providers were also available and defendants remained free to choose among the various public and private sources of representation.

We hypothesize that this proliferation of possibilities for the indigent defendant would provide a much needed spur for innovation, effectiveness and loyalty to client interests in delivery of criminal defense services. The principal risk of such an approach is two-edged. Would it successfully protect the state’s legitimate interest in avoiding excessive costs, and if so would it still successfully elicit quality defense services for the poor? To explore the question whether a voucher system can avoid these twin pitfalls, we need to examine in detail the form of reimbursement that the voucher would guarantee. We consider two possibilities: lump-sum payments and variable payments based on services actually rendered.
1. Lump-sum payments

With a lump-sum voucher, the state would grant a fixed amount to cover the cost of defense but leave to the defendant the decision of what services to buy with the money provided. The transition to lump-sum vouchers would be easiest in counties that currently compensate appointed counsel or contract counsel on a lump-sum basis. As at present, the amount of the voucher would depend on the nature of the charge, with different rates for capital cases, other felonies and misdemeanors, but unlike the present system the voucher could be cashed by any provider chosen by the defendant who was legally eligible to provide the service—that is, to practice before the relevant court.

When first implemented, a voucher system of this sort would cost no more than the prior system of representation; in principle, each voucher would be worth exactly what the county had previously been paying per case for indigent defense services. We would expect that plan administrators would find it cost-effective to make the schedule of voucher payments more discriminating—for example, linking lump-sum amounts to the particular offense charged and perhaps to other observable features of the case, such as whether it is resolved by guilty plea or by trial. But initially at least, average payments per case would be no higher than before.

Over time, the voluntary choice features of a voucher system for both attorney and client would exert some upward pressure on the indigent defense budget. If the payments offered were insufficient to attract sufficient numbers of qualified attorneys, a county that had previously relied upon conscription would have to raise the amount of its vouchers. The resulting addition to the county’s budget would not, of course, represent an increase in real economic cost, but only a transfer to the public of costs that had been borne by attorneys previously conscripted at below-market rates.

In such a system, the principal constraint on providers is market pressure, working through reputation. Just as in the market for ordinary legal services, defense firms will wish to establish a reputation for effectiveness, in order to attract more clients. A lawyer might be tempted to pocket the lump-sum fee and then stint on the time he devotes to the case, but this danger already exists in the fixed-fee appointment systems that a lump-sum voucher would replace. The difference under a voucher plan is that, as in any market transaction for service at a fixed price, stinting on service risks client dissatisfaction and, through reputation, a loss of future business. There is no such prospect for preventing meager service when the flow of future clients is controlled by the county or the court.

How well reputation will work depends in part on how well informed potential clients are about the performance of firms. While the state’s primary role in such a system is providing the voucher, there is no reason why it cannot also provide information. The court or county government could inform indigent defendants as to which firms it believes do a good job. Defendants would be free to discount the recommendation if they suspected that the state was more concerned with its interests than with theirs. Such an arrangement allows defendants to have both the informational advantage of state choice of provider and the incentive advantage of defendant choice.
So long as a lump-sum voucher is set at a level sufficient to make it attractive to criminal defense practitioners, this approach provides one way to solve the incentive problem. Not only does it use consumer sovereignty to constrain the lawyer to act in the interest of his client, it also fixes the payment obligations of the state, and thus eliminates any potential for the lawyer to increase his income at taxpayers’ expense.

The lump-sum voucher has another interesting incentive characteristic. Since the amount provided will normally increase with the seriousness of the charge, the voucher model would tend to deter prosecutors from inflating the charge. A prosecutor who follows such a strategy, to bluff the defendant into pleading guilty to a lesser count, increases the resources available to the defense and thus makes conviction more difficult.

A lump-sum voucher provides the defendant insurance against type 1 risk, uncertainty about whether a particular person will be indicted and if so, for what offense. Such a voucher also insures the defendant against type 2b risk, the risk that his case will turn out to be complex after an attorney has accepted it. Such insurance is implicit in the agreement made by a provider when he accepts the case. Defendants choose providers in terms of the total package they offer, including service for both complex and simple cases. As long as the cases cannot be distinguished in advance, a provider has an incentive to offer good service on complex cases as part of a package intended to attract clients because this is the only way to get simple cases.

The most serious disadvantage of the lump-sum voucher is that it provides no insurance against type 2a uncertainty, no protection for the defendant who has a complex case identifiable as such before the lawyer accepts it. Because the provider gets a fixed payment, he will prefer, so far as possible, either to take only simple cases or to take complex cases only on the understanding that he will not try very hard to win them. One cannot solve this problem by merely requiring providers to agree, like common carriers, to accept all comers. All a firm need do to protect itself against complex cases is do an inadequate job of defending them, thus saving money and developing a reputation that will keep away future clients with complex cases. This might be a serious argument against a voucher if the current system of indigent defense provided substantial insurance against the complex-case risk. But it does not. As we have pointed out, many counties currently provide only a lump sum for indigent defense, and thus replicate the disadvantage of a voucher without its advantages. Many more jurisdictions provide a variable amount but with a low ceiling, thus in effect offering either a lump sum or only minimal insurance. For jurisdictions that currently compensate counsel by a lump-sum payment or an hourly rate with a low cap, a voucher structured in the same way would cost taxpayers no more and would leave defendants unequivocally better off.

Nonetheless, the problem of type 2b uncertainty suggests that the lump-sum voucher is far from ideal. It is therefore important to explore possible improvements upon the lump-sum voucher approach. The next section analyzes several more complicated forms of voucher payment.
2. Hourly-rate Vouchers and Other Variations

One alternative would be for the voucher to authorize payment at a predetermined rate per hour, with a firm or presumptive cap and some possibility for a court administrator to review whether the time spent on the case was reasonable. The Canadian province of Ontario has used such a model for some time, apparently with considerable success. This approach is a compromise between the lump-sum voucher (which completely disengages the state from control over an individual provider’s compensation) and the more heavily regulated approaches discussed earlier.

The hourly-rate voucher improves the system as insurance, but brings back some of the incentive problems that a lump-sum voucher is designed to avoid. If the hourly rate is compensatory, it leaves the attorney with an incentive to spend more time than necessary on the case. Government review of fee claims is therefore essential in an hourly-rate voucher plan, as it is in existing programs that compensate appointed counsel at an hourly rate. Unfortunately, from the taxpayer’s perspective government review is a costly and imperfect monitoring device, while from the defendant’s perspective supervision by a court administrator or judge reduces his freedom of choice; the court system could punish attorneys who prove too loyal to their clients and chill their willingness to serve future defendants by permitting them only minimal fees.

These drawbacks to the hourly-rate approach would count as serious defects in this sort of voucher, except that each of them is equally present in existing hourly-rate plans for appointed counsel. The voucher approach is no worse in these respects and at least has the advantage of using the defendant’s power of choice as a reason for the attorney to take her client’s interests into consideration. Once a jurisdiction has opted to compensate appointed counsel on an hourly-rate basis, there are unequivocal welfare gains in offering defendants a “portable” voucher structured in the same way.

Since an hourly rate voucher gives the taxpayer less security than a voucher for lump-sum payment, logic alone cannot dictate the choice between these two methods of structuring compensation within a voucher system. To some extent the relative merits of these alternate approaches will depend on local conditions and on the level at which lump-sum and hourly payments are set. These matters would provide fruitful areas for investigation, as would the possibility of giving defendants a choice between lump-sum and hourly-rate vouchers. Current experience suggests that hourly rates, combined with after-the-fact monitoring, lead to more responsible and effective representation, without uncalled-for demands on the state budget. And whatever the variation in how voucher plans were structured, any of the voucher models we have considered will improve upon partial deregulation plans by permitting greater diversity of organizational form for service providers and wider scope for the operation of free choice on the part of indigent clients. Voucher models offer a package of incentive, insurance and information benefits that is clearly superior to those of either partial deregulation or, a fortiori, preservation of the status quo.
D. Objections to Voucher-Based Reforms

1. Will a Voucher Approach Prove Effective in Practice?

Our primary goal in proposing a voucher approach has been to use the engine of free choice and consumer sovereignty to improve the effectiveness of indigent defense services. But several practical concerns raise questions about whether a voucher approach would really work. We examine both economic and non-economic concerns.

a. Resource levels

Until now we have put to one side the question of how generously indigent defense services will be funded; we have simply argued that, with whatever resources society allocates to indigent defense, freedom of choice will enhance the quality of the services delivered. Among those committed to the improvement of indigent defense, however, there is an understandable preoccupation with funding levels. There are legitimate concerns that without large increases in the resources devoted to indigent defense, the reforms we propose may make little difference. Conversely, if the necessary funding increases can be achieved, it might seem that the reforms we propose will no longer be necessary. We recognize that funding levels have major impact on the quality of defense services and will continue to do so under the voucher regimes we propose. Nonetheless, as we explain in this section, freedom of choice remains a critical element in a plan for achieving effective defense services, at any level of funding.

If funding levels remain low, the pool of attorneys who serve the indigent will continue to include both able, altruistic lawyers and minimally competent attorneys with few other opportunities. It might appear that our proposal to end conscription, when combined with low resource levels, would reduce the number of able attorneys serving the poor. But the only attorneys lost in this way would be those who try to minimize the time they devote to indigent defendants because they prefer not to serve. The end of conscription would not preclude able attorneys from serving at below-market rates, and in fact would help insure that those who do so are participating out of genuine altruism and concern for client interests.

Nonetheless, if resource levels are low, many minimally competent lawyers will remain in the pool of providers. In such an environment, so common in American cities today, client needs are at risk under the best of circumstances. But client choice becomes especially important when resources are meager, because it affords defendants some reason to trust their attorney, if only by providing an escape from the worst lawyers.

A voucher plan also can have major impact on the question of what the long-run resource level will be. Client choice and a bar on conscription will tend to render funding inadequacies more visible and bring more effective pressure to correct them.

If resource levels are increased, the predicament of the indigent accused will improve, but client choice will remain a crucially important safeguard. Public defenders and appointed attorneys will no longer find it impossible to devote adequate time to their cases, but apart from altruism, such attorneys will still lack an affirmative incentive to do the best
job for their clients. In fact, if indigent appointments become lucrative and public defender salaries are raised, the attorneys involved will have even stronger reasons than at present to curry favor with court officials upon whom their positions depend. So client choice will remain essential, even with ample funding, to assure that attorneys focus on satisfying clients rather than the court.

b. Non-economic concerns

A non-economic element also affects prospects for a voucher system. What risks do we run in making the profit motive a more prominent factor in indigent defense practice? At present, idealism attracts many able lawyers to serve the poor, and these attorneys provide one of the few bright spots in the otherwise dismal picture of American indigent defense systems. In a more profit-oriented atmosphere, would fewer lawyers be drawn to this kind of work? Would attorneys in profit-oriented firms lose their idealism? Parallel concerns arise with many other proposals to substitute market arrangements for various forms of public service.

These risks should not be taken lightly, especially in an area where, as in indigent defense, idealism has played a vital role. The structure of a voucher model suggests one answer to the problem. Voluntary arrangements and free choice do not mandate a preoccupation with profit. Bar leaders could still form nonprofit corporations and hire idealistic lawyers on salary, just as happens now in Community Defender Associations. Defenders organized as government agencies could likewise emphasize public service in their recruiting and daily operations. Such organizations should have no difficulty attracting clients (and vouchers) if their performance lives up to their ideals. And if altruism permitted such firms to hire attorneys at below-market rates, they would have a further advantage that should translate into larger staffs, lower caseload ratios and more support services than profit-oriented firms could provide. The market approach we urge in this paper is not inconsistent with preserving what is best in existing systems for indigent defense. Indeed, we believe flexibility and client control are essential to insure that the good qualities in current indigent defense practice will be protected and encouraged.

2. Are Improvements in Indigent Defense Socially Desirable?

In arguing for freedom of choice and a system of vouchers as a way to improve the quality of defense services, we have taken for granted that such improvements would be a good thing. But some scholars and a substantial portion of the general public may disagree. That disagreement, though seldom openly articulated, may play a large behind-the-scenes role in explaining resistance to improving indigent defense. We believe it useful to try to make explicit the reasons for that resistance and our response to them.

One source of skepticism about the value of an effective defense is a widespread view about the way that an effective lawyer can help his client. Do the special skills of the high-priced lawyer typically serve to demonstrate the innocence of someone who was falsely charged, or do they more often enable a guilty person to get off on a “technicality”? Much of the resistance to providing better indigent defense no doubt
reflects the latter view. If that view is correct, then an improvement in the quality of defense services may occasionally keep an innocent from being convicted, but its main effect will be to make conviction of the guilty more difficult, thus lowering conviction rates, reducing the deterrent effect of criminal punishment and increasing the amount of crime.

We do not know of any way to establish whether effective lawyers help the guilty more often than they help the innocent. But even if that pessimistic view is empirically correct, it represents an obvious normative mistake. Rules of criminal procedure that permit the guilty to escape on mere “technicalities” may need to be reconsidered on their merits, but there is no justification for undermining those rules covertly by making them hard for one subset of defendants, the indigent, to invoke. So long as such rules remain on the books, they reflect presumptively legitimate goals, whether related to or distinct from protection of the innocent, and counsel should be able to invoke those rules effectively in order to promote the social values they serve.

A related but even broader claim is that virtually all defendants presently convicted by our criminal justice system are in fact guilty, so that improvement in the quality of indigent defense is unimportant. This view rests, again, on an unprovable empirical premise; recurrent news stories about demonstrably innocent men convicted of serious crimes require one to approach this claim with skepticism. Nonetheless, some scholars have sought to provide theoretical support for the view that nearly all convicted defendants are guilty. Judge Richard Posner, for example, has argued that police and prosecutors, faced with tight budgets and high crime rates, have enough to do convicting the guilty and are therefore unlikely to waste scarce resources trying to convict the innocent.

We find arguments of this sort unconvincing on several grounds. First, even in a society of overworked police and scarce prosecutorial resources, there are several sources of serious risk to the innocent. Even if prosecutors consistently select only their easiest cases, there is no guarantee that ease of conviction will correlate closely with actual guilt—especially for poorly represented defendants. Indeed, high crime rates, scarce resources and a weak system of defense may drive prosecutors to seek an easy conviction of the first suspect at hand, rather than pursuing a more thorough investigation that might exonerate him.

A second problem with the hypothesis of prosecutorial triage is more basic. To posit that prosecutors consistently select only “easy” cases seems to beg an important part of the question. If we assume, with Judge Posner, that police and prosecutors screen out weak cases to maximize the value of scarce resources, we are positing a particular kind of rational behavior and putting aside both irrational behavior such as vindictiveness, and any rational but self-seeking behavior that does not require the socially, optimal use of scarce resources. If those sorts of executive misfeasance seldom or never occur, effective judicial oversight and other checks and balances obviously become superfluous. But such a restricted behavioral model, however useful for some analytic purposes, simply assumes away the very reason why we have courts and defense counsel in the first place.
In addition to making it less likely that innocent defendants will be convicted, an improvement in the quality of defense services has other desirable effects. One is to reduce the injury the legal system does to innocent defendants who are eventually acquitted, but would have been released sooner and at lower cost to themselves if they had been adequately represented. A second effect is to provide more complete information at sentencing and thus to make it more likely that judges will impose socially appropriate punishments on the guilty.

We recognize that improvements in indigent defense, however desirable, cannot be pursued indefinitely, regardless of cost. But since a voucher system can be instituted with whatever resources a state decides to allocate to defense services, the argument against our proposals is in effect an argument that improvements in indigent defense are undesirable even if they entail no additional cost. That argument is simply an argument against the adversary system. It implies that lawyers who try hardest to get their clients acquitted are, on net, an obstacle to justice, even when they are doing their job with very limited resources. This perspective strikes at the heart of our system of criminal justice. It is of interest, in part, because it draws attention to the degree to which our present system may have become, at least for indigent defendants, inquisitorial in substance even if adversarial in form.

IV. CONCLUSION

We can see only two grounds, other than inertia, on which a reasonable person might defend existing institutions for defense of the indigent. One is the belief that defense lawyers are so bound by their professional ethics that they will consistently sacrifice their own interest to the interest of clients to whom they are assigned. Another, and less optimistic, belief is that almost all indigent defendants are guilty, if not of the offense charged than of something else, and that the real business of the court system is the administrative task of allocating punishments while maintaining a polite fiction of concern for defendants’ rights.

These arguments are both unconvincing and inconsistent with the underlying premises of an adversary system of justice. We conclude that present institutions for providing criminal defense ought to be replaced with a voucher system, in order to provide indigent defendants with freedom of choice and to provide their attorneys with the same incentive to serve their clients that attorneys have always had when they represent clients other than the poor.
§ 9. The Privatization of Force

§ 9.1. General


INTRODUCTION

Should our punishment, policing, and military institutions be public, private, or both? Is there a special link between the project of government and the exercise of force? . . .

In the past three decades, private punishment, policing, and military markets have blossomed and boomed in liberal states. Private prisons, police, and armies have been popping up around the world, punishing criminals, fighting crimes, keeping peace, and waging war. The use of force has generated unprecedented profits, and the boundaries between public and private uses of force have become increasingly blurred. Observers of these trends expect them to continue and accelerate. This Article brings these three trends under one rubric: the privatization of force.

Lately, the privatization of force has stirred up some big news in liberal states. In the United States, the Supreme Court has issued two rulings on the liability of private prisons, Congress has eliminated private baggage and passenger screening in airports, and the Department of Defense has employed private security firms to protect the Coalition Provisional Authority in postwar Iraq. . . .

. . . In separate disciplines, there have been vibrant public policy debates over the privatization of punishment, policing, and military force. In this Article I collect and critique these public policy debates. I identify the most essential themes in each argument and discover that each one fails to engage our most basic ideals. When we talk about the privatization of prisons, police, and armies, we talk in the same generic, mundane terms that we use to describe the privatization of schools, hospitals, welfare systems, and the like. By resorting to clichés, we have missed the special problems presented by the proliferation of private prisons, police, and armies. As a result, we have failed to articulate any fundamental critiques.

The time is ripe to assume a more profound perspective vis-à-vis the privatization of force. Private armies, police, and prisons are not only objects of public policy to which we can apply our customary arguments from the left and the right; they are also challenges to one of our most fundamental, venerated axioms of liberal thought—the idea that “the state” has, must have, or should have a “monopoly of force.” Unlike past scholars, I take the privatization of force as an invitation to revisit an old feud over the appropriate relationship between government and force. Should the liberal state have a “monopoly of force”? If so, what should such a monopoly state look like, in institutional
terms? These are the two central questions that this Article raises and answers via quirky couplings of economic and political concepts.

The Article proceeds in five parts. Parts I and II hang the theoretical and factual backdrop for my analysis. Part I recounts the story of the state of nature and defines a few basic terms: “liberal,” “force,” “public,” and “private.” This Part sets forth the monopoly thesis: the liberal idea that the state has, must have, or should have a monopoly of force. Part II provides a brief history of force. It gives a quick glimpse of the earlier monopolization and recent privatization of punishment, policing, and military force in liberal states.

Taken together, Parts I and II portray an apparent tension between the theory and practice of liberalism. On the one hand, we have the idea that force must be monopolized; on the other hand, we have the fact that force has been privatized. How might this tension be resolved? Should we monopolize or privatize force? Is the answer the same for punishment, policing, and military force? Or is the tension between theory and practice only apparent? Are our choices between public and private more complex? When we consider the privatization of punishment, policing, and military force, what principles should guide our thought?

The rest of the Article speaks to these issues. Part III introduces the customary four-box matrix, which conveys a distinction between two important concepts: the “demand” for force and the “supply” of force. In the rest of the Article, the supply/demand matrix serves as our polestar, guiding us toward fundamental critiques.

First, in Part IV, the matrix helps us navigate the maze of our public policy debates. These debates raise many different principles and concerns: the values of (1) economic efficiency, (2) egalitarian distributions, (3) public goods, (4) public accountability, and (5) human rights, and the dangers of (6) industrial influence, (7) market failure, and (8) cultural commodification. When we apply the supply/demand matrix to each of these arguments, we slice through the polemical morass. We perceive quickly that our public policy arguments have generated powerful critiques of the private demand for force but weak critiques of the private supply of force. We see that our current approach is too generic, and as a result, it is too superficial. We have yet to identify any fundamental principles that constrain the privatization of punishment, policing, and military force.

In Part V, I restart my analysis from scratch. I return to the state of nature, the intellectual foundation of liberal thought, in search of such principles. I retrace our first steps from anarchy to monopoly to state at a more deliberate pace. Step by step, I construct a new version of the state-of-nature story and a more fleshed-out picture of the monopoly thesis. I dub my vision the “liberal theory of force.”

My argument proceeds in four stages. The first three are supply-side stages. First, I introduce an “ultraminimal state” in which the demand for force is completely public and the supply of force is completely private. Next, I introduce a “military state” in which the demand for force is still public but the supply of military force is now public too. (The supply of policing and punishment remain private in the military state.) I argue that the
military state is more stable than the ultraminimal state because it is less vulnerable to the “problem of loyalty”—the danger of military revolt.

In stage three, I generalize the argument. I develop the concept of the “intensity of force,” and argue that the intensity of force justifies the state’s monopoly of the supply of force, in equal proportions, to equal degrees. It turns out to be a surprisingly blunt line of thought that remained overlooked. I argue that punishment is the least intense form of force, policing is the median form of force, and military force is the most intense form of force. In less opaque terms, armies are more powerful than prisons. For this simple reason, I argue that the liberal state must monopolize the supply of military force but need not monopolize the supply of punishment. In an ideal liberal state, armies must be public, but prisons may be private. Police, I conclude, are a more mixed and complex case: They fall between the two extremes of armies and prisons. In the fourth stage of my argument, I include a few brief remarks on the private demand for punishment, policing, and military force.

This Article has several aims, some more ambitious than others. If I am fortunate, I will convince you that my blend of liberal theory is both distinctive and worthwhile. If not, I hope at least to persuade you that the privatization of prisons, police, and armies tells us something important about ourselves: that we do not yet understand the liberal state’s raison d’être, that we have left stones unturned amid the foundations of liberal thought. On a more practical level, I aim to give you a helpful framework for thinking about whether our punishment, policing, and military institutions should be public, private, or both.

I. ANARCHY, MONOPOLY, AND STATE

A. The Story of the State of Nature

. . . We start with a story: Once upon a time, there were individuals who lived in a state of nature, a community without government. These individuals thought that they had things called “rights”—or at least, they thought that they should. But they had a problem: On the whole, they were a rational, self-interested lot, and resources were scarce. Human nature and mother nature worked against one another, making life rather unpleasant. The individuals frequently fought over resources and violated each other’s rights. In other words, they got away with whatever they could, and in the absence of government, they could get away with a lot.

This tale is very old and often told, so you might remember its happy ending: The individuals eventually become so fed up with fighting against each other that they start working together. One way or another, they end up creating an organization that determines and enforces rights, or prevents and settles disputes. In some accounts, this organization begins as a kind of a “market” or “corporation” that is structured around a voluntary system of prices. (You pay only for the protection you want; you get only the protection for which you pay). In the end, though, it always ends as a “state” or “government” that is organized around a mandatory system of taxes. (Everyone pays for
protection; everyone is protected.) Out of private competition emerges public coercion. Without further ado, then allow me to introduce the “liberal state.”

This state-of-nature story is the bible of liberalism. It lays down the deepest dogmas of liberal thought and spins them into a forceful account of our social, economic, and political origins. For at least four centuries, this tale of genesis has been passed down and retold by a long line of liberal philosophers, politicians, and lawyers, among others.

Of course, it has been told in many different ways—as many ways as there have been liberal theorists, or at least, liberal theories. In each rendition, some basic concept like “nature,” “reason,” “freedom,” “justice,” “democracy,” or even “efficiency” plays the lead role, motivating and organizing the individuals in the story and justifying the state they create. In the most expansive accounts, the liberal state does a good deal more than just prevent and settle disputes. It also pursues many other public projects—from transportation to education, from health care to welfare. But even in the most minimal accounts, the liberal state encodes rights into laws and uses threats and acts of physical coercion to enforce them. It polices and punishes; it protects and defends. Thus, no matter who tells the tale, or how it is told, the state-of-nature story always ends with the same basic punchline: The state has, must have, or should have a monopoly of force.

Let us call this conclusion the “monopoly thesis.” . . . Today it is typically treated as a truism, a self-evident definition or principle of government, a natural and necessary fact of life. . . .

**B. The Monopoly Thesis**

So liberal states “monopolize” force. They did. They do. They must. They should. Huh. As story endings go, this sounds like a good one, but what does it mean, and why is it so? These are the central questions of this Article, and the answers turn out to be surprisingly complex. But we should not get ahead of ourselves. We should start by noting a few things that the monopoly thesis does not say by clarifying a few basic terms.

1. **Liberal**

First, “liberal.” In this Article, when I say that something is “liberal”—e.g., a philosopher, a state, or an idea—I do not mean to say that it is left-wing or socialist rather than conservative, right-wing, or capitalist. Everything on that list is “liberal” for my purposes. Instead, when I say that something is liberal, I mean to say that it follows the philosophical, political, and legal tradition of “liberalism,” which has strongly influenced and reflected the development of modern, industrialized states in the West. What is liberalism? In this Article, liberalism includes any set of ideas based upon the state-of-nature story itself. So the first thing to remember is that the monopoly thesis is a liberal thesis. It claims that liberal states must monopolize force.

2. **Force**

Second, “force.” Force is any legitimimized threat or act of physical coercion. Note four things: First, force is only physical coercion, or violence; it is not social, economic, or
any other kind of coercion. Second, force is only legitimized violence; it is not crime, terrorism, or other kinds of violence that are not legitimized by liberal states. Third, force includes threats as well as acts. Fourth, for our purposes, force usually refers to only specialized, organized kinds of legitimized violence—that is, the force exercised by modern punishment, policing, and military institutions. (There is a brief moment when it also includes individualized threats and acts of force, like self-defense and self-help, but when it does, I make that clear.) Taken together, these points suggest the second thing to remember about the monopoly thesis: Liberal states must monopolize punishment, policing, and military force.

3. Public

Third, I will address “public” and the related term “monopolization.” These terms are a bit more complex so a basic framework is helpful. In the modern liberal world, we often use three basic metaphors to describe the kinds of organizations through which we recognize and respond to social problems: “community,” “market,” and “state.” I cannot spend much space on these metaphors here, so I will just say a few things.

Communities shape behavior through duties: “I ought to do this for you because it is the right thing to do.” They operate through overlapping cultural systems of beliefs, symbols, and rituals, such as morality, religion, family, friendship, work, voluntarism, and neighborhood.

Markets shape behavior through profits: “I want to do this for you because I will receive something of value for doing it.” Markets operate through economic systems of reciprocal, self-interested exchange in which prices signal how much supply and demand there are for particular goods and services.

States shape behavior through force: “I must do this for you or else I will lose something of value for failing to do it.” States operate through systems of political and legal authority in which principles and policies are developed, interpreted, and applied by legislative, adjudicative, and administrative institutions. In addition, states operate through systems of political and legal power in which principles and policies are enforced by punishment, policing, and military institutions.

Of course, there have never been any pure communities, markets, or states in real life. These words signify tendencies, principles, and structures rather than actual institutions. . . Thus, when we say that a “community,” “market,” or “state” is doing something—or, for that matter, when a “corporation” or “government” does something—we do not really mean it. We are using a metaphor. We actually mean that people are doing something, and that on balance, they tend to use particular means and act with particular motives.

Now we can make a first pass at the monopoly thesis and the difference between “public” and “private” punishment, policing, and military institutions. The monopoly thesis says that liberal states must “monopolize” force. This is obviously a vague statement which could mean several different things, some of which are more plausible than others. We will consider each possibility in due course.
But for now, let us start with the most obvious one: On its face, the monopoly thesis seems to suggest that punishment, policing, and military institutions must be “public”—that punishment, policing, and military tasks must be performed primarily by states rather than markets or communities. More broadly, it suggests that the people who exercise force must be motivated and organized primarily by political and legal authorities rather than economic exchanges or cultural duties. As I have already mentioned, this is a widely held thesis. It should sound intuitive and sensible—perhaps even inevitable—to our modern, liberal, American ears.

4. Private

Yet I also suggested that these days it has become a less plausible thought. Over the last thirty years, there has been a remarkable privatization of punishment, policing, and military force in liberal states. . . . Our punishment, policing, and military institutions are becoming less like states and more like markets—more motivated by profit, more shaped by economic exchange.

II. OUT OF ANARCHY? A BRIEF HISTORY OF FORCE

. . . Mind the gap: On the one hand, we have an old set of theories that (apparently) require the monopolization of force. On the other hand, we have a new set of practices that (apparently) constitute the privatization of force. Questions leap out: How and why did this rift emerge? Is the monopoly thesis fantasy or fact? In our history, have so-called liberal states ever actually established, or even sought to establish, public monopolies of punishment, policing, or military force? If so, do liberal states still maintain such monopolies now?

In this Part, I take up these questions. If they seem like obvious questions to ask, it might surprise you to learn how rarely liberal philosophers have bothered to ask them. . . . They have presumed that liberal states monopolize force, but they have rarely troubled themselves to examine, explain, or prove the point.

A. The Monopolization of Force

If we take a very broad view of human history, it is easy to see why these questions have been so widely ignored. In contrast to ancient and medieval civilizations, today’s world is clearly distinguished by a greater separation of public and private domains. This is probably true throughout the West and certainly true in those Western states that call themselves “liberal.”

There is no question that in the Europe of 1000 A.D., for example, there were no distinctly “public” punishment, policing, and military institutions at all. In fact, there were not even centralized, national states, nor were there many institutions that specialized in the exercise of “punishment,” “policing,” or “military” force. Today, Western states have all of these institutions, and they are numerous and powerful indeed. Thus, it seems safe to say that in the long run, the monopoly thesis rings more or less true: In relative terms, today’s “liberal” states more or less “monopolize” force.
Presumably, this explains why so many people assume that they did, they do, they should, and they must.

But one thousand years is a rather long run. When we take such a broad view of history, we obscure a great deal. We fool ourselves into thinking of the emergence of public punishment, policing, and military institutions as a necessary, deliberate, one-way shift from “private” to “public,” which coincided roughly with the onset of things like “modernity,” “industrialization,” “centralization,” “bureaucratization,” “liberalization,” and “democratization.” It was not and did not. The trend cannot be categorized in such simple, developmental terms. At times, the establishment of public punishment, policing, and military institutions was surprisingly contingent, reluctant, irregular, and late.

In fact, throughout the millennium, states periodically resisted the monopoly logic by making punishment, policing, and military force more private in many respects. As any historian or sociologist knows, examples abound: European privateers, mercenaries, and mercantile companies; English thief-takers; American Pinkertons; and convict leasing systems in Europe and the United States. Each of these examples provides important reminders that in the grand scheme of history, today’s private punishment, policing, and military markets are not wholly new.

But in the early twentieth century, they seemed to be things of the past. By then most liberal states had more or less achieved monopolies of punishment, policing, and military force. . . .

B. The Privatization of Force

About thirty years ago, however, something strange began to happen. After almost a half century of exile, the prodigal market returned.

[Rosky describes the development of the private policing, prison, and military industries. See § 0 above.]

4. Two Observations

. . . [W]e can make two obvious points. First, we have not come full circle. Liberal states are not back where they started one hundred [or] one thousand years ago. Today’s private punishment, policing, and military corporations are hardly comparable to the practices of convict leasing, thief-taking, strike-busting, colonization, mercenarism, or privateering that prevailed in centuries past. . . . [P]rivate punishment, policing, and military corporations . . . still live largely in the shadows of exceptionally powerful governments. In many cases, they exist largely at the request or behest of such governments.

For this reason, we should be suspicious of analogies between past and present examples of the privatization of force. Such analogies tend to obscure one of the greatest political, economic, and social transformations of recent history: the rise of modern, liberal administrative states. In today’s world, liberal states have many more political, economic, and cultural resources available to hold modern punishment, policing, and military markets in check. How liberal states should use these resources is . . . less obvious . . . .
The second point is the flipside of the first: Although the privatization of force has just recently begun, it constitutes a striking, decisive departure from our recent past. Before this trend began, most liberal states had been monopolizing punishment, policing, and military force for at least fifty years, and in some respects, for nearly three centuries. Thirty years ago, private armies and private police were nearly forgotten, and private prisons did not exist. Today, these corporations are doing a thriving business around the world . . . . I have highlighted the most extreme examples because they drive my point home: Over the last thirty years, liberal states have started to turn away from—or at least reconsider—the old adage that the state must monopolize force.

III. THE MATRIX: SUPPLY AND DEMAND

. . . Before we delve into our public policy debates, we must first introduce a fundamental distinction: the distinction between the supply of and the demand for force. It is a complex and broad distinction that must be deliberately defined. . . . The distinction will deepen our understanding of many relevant topics: our recent privatization of punishment, policing, and military force; our public policy debates over these trends; and our abiding attachment to the old idea that liberal states must monopolize force.

In this brief Part, I spell out the distinction between supply and demand from two different perspectives: at first through an economic lens, and then through a political/legal lens. In economic terms, the supply/demand distinction represents distinction between the “provision” of force and the “payment” for force. In political/legal terms, it represents the distinction between the “exercise” and “allocation” of force. But these are opaque terms that must be explained.

There are actually two distinctions here: between “public” and “private” and between “supply” and “demand.” [T]he distinction between “public” and “private” modes of organization can be broken down into two component distinctions: First, there are public and private ways to create supply; second, there are public and private ways to express and control demand. This suggests the following two-by-two matrix . . . :

![Figure 1](attachment:figure1.png)

. . . First: Please remember that “public” and “private” are always metaphors and matters of degree. . . . Second: In the United States, roughly one half of . . . tax revenues . . . is spent in the lower left box of the matrix, to pay for privately supplied goods and services. That fact has very little to do with the privatization of punishment, policing, and military
force, but it provides a sense of how widespread the “public demand, private supply” arrangement has become in today’s liberal states. Lastly, and most importantly, please take a moment to put your mind into the matrix. It is exceptionally abstract, and rather opaque, but worth a bit of intellectual struggle.

A. Provision and Payment: The Economics of Supply and Demand

. . . What do I mean by “public” and “private” ways to “create supply” and “express or control demand”? In economic terms, I mean public and private modes of provision and payment. . . . Provision represents the following question: Who provides punishment, policing, and military force—governments or corporations? If you are not an economist, that question might sound awkward. In everyday life, we do not normally talk about “supplying” or “providing” punishment, policing, and military force. But that is precisely the point: The economic perspective encourages us to think of punishment, policing, and military force as “services” that governments and corporations “provide.”

. . . Payment represents the following question: Who pays for punishment, policing, and military force—governments or “customers”? . . . Instead of payment, we might also say “spending.” When customers pay for something, we often call that “purchasing” or “consumption”; when governments pay for something, we often call that “financing.” . . . For the purposes of this Article, “payments” are not only payoffs, or monetary exchanges. They are payoffs attached to economic decisions. They are private and public arrangements to purchase and finance the provision of force.

. . . [T]he choices between public and private provision and payment are not black and white. . . . “[P]rovision” can be broken down into three distinctions between public and private forms: ownership, management, and employment. In contrast, it is not very helpful to break down “payment” into separate parts. But it is helpful to think about the difference between public and private payment as the difference between taxes and prices: Public payments are relatively collective and mandatory, like most taxes; private payments are relatively individual and voluntary, like most prices. With this more economic perspective in mind, let us revisit the matrix:

FIGURE 2

<table>
<thead>
<tr>
<th>Provision</th>
<th>Public</th>
<th>Private</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>U.S. Armed Forces</td>
<td>U.S. Postal Service</td>
</tr>
<tr>
<td>Private</td>
<td>Private prisons</td>
<td>Private police</td>
</tr>
</tbody>
</table>

619

Draft — do not distribute!
Hopefully, this makes things a bit clearer. Rather than running through each box, I will compare two examples: private prisons and private police. In terms of provision, the two examples are similar: Corporations provide both private prisons and private policing. But in terms of payment, the two examples are different: Customers pay for private policing, but they do not pay for private prisons. Only governments pay for private prisons.

The difference between private punishment and private policing is even more complicated. In my example, I included only private police hired by private customers, such as individuals, associations, and corporations. In other words, I excluded private police hired by governments. But in the real world, governments often enter the private policing market as “customers” (or “consumers”) and pay for private police. In the United States, federal, state, and local governments spend almost as much on private police as they spend on public police. In those cases, private police are roughly like private prisons: provision is private, payment is public. Thus, in the real world, “private policing” appears in both the lower left and lower right boxes.

As this point suggests, it is possible to think about these examples and boxes in much broader terms. For instance, imagine that “public police” and “public prisons” were included in the matrix as well. They would both appear in the upper left box, alongside the . . . Armed Forces. Including these two examples reveals that we can speak of the overall provision of and payment for “policing” or “punishment.” . . . From this broad perspective, we would say that the provision of policing is both public and private, or “mixed,” and the provision of punishment is also mixed. The payment for policing is also mixed. But the payment for punishment is different: It is purely public. Only governments finance punishment. This is a special quality of punishment: You cannot “shop” for imprisonment services. . . .

. . . [L]et me clarify two further points: First, you might have noticed that PMCs do not appear on the matrix. In the private military market, there are some confusing details. But on the whole . . . , we can say that private armies that are owned, managed, and staffed within the territories and jurisdictions of liberal states are roughly like private police who perform different tasks, and work almost exclusively for government clients. By definition, they are privately provided, but in most cases, they are publicly financed.

[S]econd . . . I am not a conspiracy theorist. I do not believe that the U.S. Postal Service has anything to do with the supply of or demand for force. I include the U.S. Postal Service just to round out the matrix: The federal government provides postal services, but customers pay for them, by paying for stamps. There are more relevant examples that fall into this box, but they are much more obscure. The clearest one is public police departments “moonlighting” for private clients in exchange for private funds.

B. Exercise and Allocation: The Politics and Law of Supply and Demand

Enough economics. Now let us shift into political/legal gears. What do I mean by “public and private ways” to “create supply” and “express or control demand”? In political and legal terms, I mean public and private modes of exercise and allocation. “Exercise” is a fairly simple concept: It is the execution, performance, or implementation of punishment, policing, and military tasks. It asks us: Who exercises punishment, policing, and military
force—corporations or governments? There is, however, one important wrinkle. From now on, we must distinguish between the “use” and “exercise” of force. The “use” of force includes only threats and acts of legitimate physical coercion. The “exercise” of force includes these threats and acts, but it also includes many supply-side decisions regarding who may use force, as well as when, where, how, how much, and against whom force may be used. Instead of exercise, we might also say “enforcement.”

The concept of “allocation” is much more complex. First and foremost, “allocation” includes all of the demand-side decisions that governments and customers make regarding who may use force, as well as when, where, how, how much, and against whom force may be used. But it also includes demand-side activities: discussions, negotiations, and inspections. It asks us: Who allocates punishment, policing, and military force—governments or customers? Instead of allocation, we might also say, “authorization,” or “legitimation.”

Like the concept of payment, the concept of allocation cannot be usefully broken down into component parts. But it is helpful to think about . . . the difference between public and private allocation as the difference between regulations and contracts: From the perspective of citizens, public allocations are relatively collective and mandatory, like most regulations; from the perspective of customers, private allocations are relatively individual and voluntary, like most contracts. We classify regulations by the public institutions that issue them: “legislative,” “judicial,” and “administrative.” Punishment offers good examples of regulations: Legislators define crimes and . . . penalties; judges fix sentences; politicians and administrators review requests for clemency and parole.

Like the taxes/prices distinction, the regulation/contract distinction should not be taken too literally. It obscures a very important exception: Sometimes governments act like customers. They sign contracts with private prisons, police, and armies. [T]hese . . . “contracts” [are also] public allocations of force. From the perspective of suppliers, they are individual and voluntary; [to] citizens, they are collective and mandatory.

One . . . key element[] in the exercise/allocation distinction is the distinction between supply-side and demand-side decisions regarding the use of force. . . . Supply-side decisions are the kinds of decisions made by managers and employees of public punishment, policing, and military institutions; and owners, managers, and employees of private punishment, policing, and military institutions. . . . Demand-side decisions are regulations and contracts . . . made by legislators, judges, administrators, and customers.

“Allocation” is easily the broadest and least intuitive concept . . . . We do not often lump together public and private allocations of force. But [we can] think of them both as members of a particular species of high-level decision regarding the who, when, where, how, and how much of force. In this sense, public and private contracts function much like sentences, warrants, and declarations of war, as well as statutes that authorize and govern exercises of punishment, policing, and military force[, all of which] allocate force.

With this political/legal perspective in mind, let us return one last time to the matrix:
Because this . . . matrix includes the term “allocation,” it highlights the point that I made earlier—that the distinction between public and private is not black and white. Look at prisons: The matrix suggests that the allocation of imprisonment is “public”—that only governments allocate imprisonment. But there are counterexamples: . . . [P]rivate persons make demand-side decisions that help determine who is imprisoned, for how long they are imprisoned, and for what they are imprisoned[; by deciding] whether and how to gather evidence, press charges, testify, speak at parole hearings, and so on. But on the whole, . . . [i]t is most often . . . governments that allocate imprisonment.

To see the same point from the other side, consider the example of the private police. The matrix suggests that the allocation of private policing is “private”—that only customers allocate private policing force. Again, there are counterexamples: Even when private police are privately hired, governments still allocate private policing to some degree. After all . . . , governments regulate private policing; they enforce contracts, civil codes, and criminal statutes against private policing firms and employees. But again, on balance . . . , [i]t is most often and most importantly customers who allocate private policing.

C. Into the Matrix

. . . I want . . . you to think of the last two distinctions—provision/payment and exercise/allocation—as . . . versions of the first and most basic distinction between supply and demand. . . . [T]he distinctions between public/private “provision” and public/private “exercise” are actually the same . . . , as are the distinctions between public/private “payment” and public/private “allocation.” It is . . . easier to convey in analogies:

Supply Analogies
1. Public Supply = Public Provision = Public Exercise
2. Private Supply = Private Provision = Private Exercise

Demand Analogies
1. Public Demand = Public Payment = Public Allocation
2. Private Demand = Private Payment = Private Allocation
I am not suggesting that “payments” and “allocations” are the same things, or that every decision to pay for something is also a decision to allocate something. Sometimes they are; sometimes they are not. For instance, if I hire a bodyguard, my payment for force would be very closely linked to (perhaps indistinguishable from) my allocation of force. But when judges send criminals to prison, they probably do not consider the economic costs of imprisonment, and they certainly do not make public payoffs to prisons. Nor should they. Sentences are public allocations, not public payments.

But the matrix is not about individual cases, nor even categories of allocations and payments. It is about the distinction between public and private organizational modes. Thus, when I draw analogies between payments and allocations, I suggest only that private allocations require private payments and that public allocations require public payments. Conversely, I suggest that private allocations do not require public payments and that public allocations do not require private payments. In modern organizations, the two pairs almost always appear together; they almost never cross-pollinate . . . :

Public Organizational Modes

1. Public Demand = Public Payment + Public Allocation
2. Public Supply = Public Provision + Public Exercise

Private Organizational Modes

1. Private Demand = Private Payment + Private Allocation
2. Private Supply = Private Provision + Private Exercise

Imprisonment provides a simple example. When (public) judges send criminals to prisons, (public) legislators and (public) administrators consider the economic costs of imprisonment, and make public payoffs to prisons. Individual victims of crimes . . . do not foot the bill . . . . When governments allocate punishment, policing, and military force, governments consider the economic costs . . . and . . . make payoffs to fund them. Public allocations require public payments. It is not a controversial thought . . .

IV. THE PUBLIC POLICY DEBATES

In this Part, I glean the most common themes . . . from our public policy debates. . . . I cannot provide a complete catalog of . . . policy arguments . . . . Instead, I offer a whirlwind tour—an index of straw arguments, a review of essential themes . . .

For each argument, my analysis follows regular techniques. First, I lay out the argument itself . . . Then I respond. In each response, I probe the strengths and weakness of the argument . . . . I explore what the argument might and might not achieve, in its most persuasive forms; more importantly, I explore what it may never do in any event.

I conclude that our . . . policy debates . . . identify . . . practical pitfalls of privatizing prisons, police, and armies, but they do not articulate any fundamental objections to the privatization of force per se. They teach us how to regulate . . . the privatization of force, but they do not justify prohibiting private prisons, police, and armies themselves.
[This is because] our public policy critiques . . . provide strong objections against the private demand for force but weak objections against the private supply of force. . . . [T]hey . . . fail to confront . . . [t]he model in which the supply of force is private but the demand for force is public. In the next Part, I pick up where our public policy debates have left off. Going back to the state of nature, I explore whether, why, and how the model of private supply and public demand would fail.

A. Economic Efficiency

1. Argument

Like all proponents of privatization, advocates for the privatization of punishment, policing, and military force make arguments motivated by economic efficiency. These arguments . . . look like this: Private companies are more efficient than governments. Private companies provide higher quality services at lower prices. They meet demands faster, more flexibly, more accurately, more completely, at lower prices. Everybody wins: Streets, prisons, and nations are safer; taxes are lower, and shareholders are richer.

More specifically, advocates say . . . : Private prisons are financed, constructed, and operational quicker than public prisons. . . [L]iving conditions are better; inmates are safer; escapes are less frequent. Most importantly . . . , they are less expensive to run . . . . Private police make preventative patrols that public police cannot afford to make. They deter crime before it is committed . . . They spend less time doing paperwork and eating donuts. They internalize the costs of policing by charging clients directly. Again, they cost less . . . . Private armies deliver expertise, training, weapons, transportation, intelligence, and soldiers quicker than governments. Private armies mobilize, fight, and win faster than public armies—and above all, they work for less.

2. Response

Or so the logic of efficiency suggests. It is deeply controversial. In public policy circles, it is easily the most hotly debated and researched issue of all. The efficiency argument has been most intense in the context of punishment, where scholars have conducted many detailed, empirical studies comparing the quality and cost of public and private prisons. Unsurprisingly, the results of these studies have been mixed . . . . These arguments often boil down to . . . which camp bears the burden of proof[.]

This quandary . . . suggests deeper, more philosophical issues. In the interest of advancing the privatization conversation . . . , I concede the . . . efficiency argument . . . . [L]et us assume that corporations provide punishment, policing, and military services more efficiently than governments. . . . [B]ut there must be more to life.

What more? By asking this question, we begin a liberal investigation of the privatization of force. . . . [C]ritics of privatization have taken several liberal turns, pressing the public policy debates beyond the economic principles of quality and cost. They have sounded their critiques in seven classic liberal themes: the values of (1) egalitarian distribution; (2) public goods; (3) human rights; and (4) public accountability; and the dangers of (5) industrial influence; (6) market failure; and (7) cultural commodification.
B. Egalitarian Distribution

1. Argument

First, critics object that private markets provide the greatest security to the highest bidders rather than protecting all citizens equally. This distributive argument appears most frequently in criticisms of private policing and military markets. It comes in two forms: static and dynamic. Both versions proceed from the presumption that policing and military security are the kinds of things that every individual and organization wants and deserves.

In the static version, critics note that private policing and military corporations sell security, whereas public policing and military institutions ration it. Thus, private security ends up in the hands of the wealthiest . . ., while public security is more evenly shared. There is an obvious rejoinder . . .: Private security markets [are] most often . . . supplements, [not] substitutes, for public security networks. Customers only purchase the services of private policing and military corporations when governments cannot or will not provide them. This response works especially well in the private military markets, in which poor states typically hire PMCs because rich states refuse to help them. . . . [I]n the real world, the only alternative to PMCs is not public armies, but no armies.

This response provokes the subtler, dynamic version . . .: [O]ur increasing reliance on private policing and military markets encourages wealthy individuals, corporations, and governments to “exit” from public security networks, leaving poor communities and states undefended. . . . [W]ealthy individuals, associations, and corporations . . . [buy] more private policing . . . [and also] pay [high] tax bills. Critics worry that eventually, wealthy customers will become fed up with “paying twice for security—once to the government, and once to hired private firms.” They will relocate to privately-owned . . . low-tax[] enclaves in which public policing levels are low. They will lobby for cutbacks in public policing institutions and demand tax vouchers for private security expenses. They will withdraw their resources from public police forces and reinvest them in private policing markets. . . . [G]overnments will [be] increasingly unable to sustain or improve public police protection [for the poor].

2. Response

This is a fine argument . . . [b]ut it does not go very far for at least two reasons. First, . . . this argument . . . does not offer any critique of the model of private supply and public demand[, such as] private prisons[, where] the demand for punishment is purely public. . . . Governments are the only customers of private prisons, so they can distribute imprisonment however they like. If governments distribute imprisonment along class lines, that is a problem created by governments . . . .

. . . Governments . . . are [also] . . . the most significant customers [in private military markets, and] are also major customers in private policing markets: In the United States, they spend almost as much on private police as they spend on public police. Thinking back to the matrix, let us call these the cases of “public demand.”
In these public demand cases, the static argument sounds hollow, and the dynamic argument makes no sense at all. The static argument supposes that when supply and demand are public—when governments provide, finance, exercise, and allocate punishment, policing, and military force—force is evenly distributed. . . . What, then, is the problem with the private supply of force? If the static argument is correct, then when governments make payments and allocations, they produce egalitarian distributions. . . .

For similar reasons, the dynamic argument does not apply to public demand. Admittedly, when governments hire private policing corporations, they spend tax dollars that would have otherwise been spent on public policing institutions. But if they distribute private security evenly, and in sufficient amounts, then who cares? Clearly, governments cannot tell themselves that they are paying “twice” for security, and they cannot withdraw resources from themselves. In both the static and dynamic senses, the public purchasing of private policing runs no risk of distributive injustice.

This leaves private demand cases, such as the private purchasing of private policing. . . . The . . . logic of [the distributive] argument[] suggests . . . prohibit[ing] . . . [private actors] from hiring private policing corporations. . . . Then they could not exit public policing institutions. They would have no other source of protection against crimes.

Of course, that is a drastic and implausible idea. It raises a whole host of special concerns about self-defense, self-help, and property rights, which I want to defer [to Part V]. . . . For now, let us focus on a more commonplace, reasonable proposal: tax and transfer. We could tax private purchases of private policing and transfer these funds to impoverished communities. Since we are concerned with the distribution of security, we could earmark these funds. For example, states could offer private security grants to public and private organizations in crime-ridden, impoverished areas or send private security vouchers or coupons to indigent citizens. With these modest reforms of private demand, some of the benefits of our burgeoning private policing markets would redound to all.

This, then, is the second conceptual limit of the distributive critiques. When they critique private demand, they are not fundamental. They are just cautionary tales, which can be answered by redistribution, one of our most standard liberal reforms.

C. Public Goods

1. Argument

Whereas the first argument is about the distribution of security, the second argument is about the quantity and quality . . . . Critics assert that under certain special conditions, private markets do not produce enough security, leaving customers and citizens exposed to crimes and attacks. The debate over the federalization (or monopolization) of airport security provides the most prominent and current example. . . . After the [9/11] attacks, many . . . argued that the privatization of airport security had pulled the industry into a “race to the bottom” in which airlines competed by keeping the quantity and quality of airport security services abysmally low.
The “quantity” and “quality” arguments seem to blend together four economic concepts: market failures, public goods, tying arrangements, and negative externalities.

In economics, “market failure” is a term of art. . . . Economists have a fairly standard list of common market failures, and everyone’s list includes the category of public goods. For our purposes, we can define goods as “public” when they are relatively nonexclusive: Once they are sold to one customer, it is too costly for the seller to stop other customers from enjoying them too. In common parlance, we call this the problem of freeriding: One person buys something, then everyone else mooches rather than buying more for themselves. In the airport security market, the logic of public goods clearly applies: Within airports—and especially within airplanes—security is not very exclusive. Once you truly protect one customer, you have already protected a whole plane of customers, if not a whole airport as well.

Thus, rather than producing goods that no one would buy, firms refuse to produce anything at all. To cure this failure, states intervene. They compel consumption. They impose taxes upon citizens and establish optimal levels of quality and quantity . . . .

There is a sense in which commercial airlines had already attempted to self-regulate—to achieve a similar result by themselves, even before [9/11]. . . . You could not buy air travel without buying airport security. Whenever you purchased a ticket, you paid for security and travel together. . . . In economic terms, this is a “tying arrangement” because the two services are bundled together.

In retrospect, this private solution was only a band-aid. . . . Price competition kept ticket prices down, and airlines kept airport security levels below par. . . . The Aviation and Transportation Security Act of 2001 . . . federalized airport security—imposing mandatory increases in levels of screening, training, staffing, and compensation.

2. Response

Before we take a closer look, this sounds like an obvious reform. Like the distributive argument, the public-goods argument is less a theory of public supply and more a theory of public demand. It is, as economists say, a theory of public “spending” or “finance.” . . . The theory suggests that when airlines purchase and allocate security, they do not purchase and allocate enough. The obvious solution, then, was not to federalize the supply of airport security. It was to federalize the demand for airport security—to put the financing and allocation of airport security under federal control. Once demand became federal, politicians could hire as much security as they desired. The problems of exclusivity would disappear.

What would this reform have looked like? Just like the actual Aviation Transportation and Security Act of 2001, a hypothetical “Public Goods Act of 2001” would have imposed a host of quality and quantity requirements. Throughout the nation’s airports, it would have increased the requisite levels of screening, training, staffing, and compensation of airport security personnel. These requirements could have been included in regulations, as the Administration proposed, but they also could have been included in
“specs,” or procurement policies, required by contracts between the federal government and private security firms. They could have been enforced through negotiations and inspections in addition to, or in lieu of, traditional enforcement mechanisms.

In some respects, this would not have been too different from the Aviation and Transportation Security Act that was actually passed. Under the authority of the Act, the FAA paid a human resources corporation, NCS Pearson, over $100 million to “recruit, test, and hire” our new federal passenger and baggage screeners. The contract required physical and psychological assessments of applicants and specified other quality and quantity goals. It is a bit strange: We federalized the supply of airport security; we even regulated the demand for airport security; yet we signed a contract with a corporation in order to carry out our own regulatory reforms.

In the end, then, the hypothetical Public Goods Act might have looked much like the real Aviation and Transportation Security Act, with one necessary distinction: the passenger and baggage screeners at airports would work for Airport Security Inc., rather than the [U.S.] Department of Justice. Is this a distinction with a difference? That remains to be seen. But this much is already clear: If there is a difference between these two solutions, it does not hinge on the nonexclusivity of airport security or the concept of public goods.

D. Public Accountability

1. Arguments

Third, critics argue that privatization diminishes the public accountability of punishment, policing, and military force. . . . [T]hese arguments draw upon two central themes: First, the distinction between public and private institutional purposes (politics vs. profits); second, the distinction between . . . accountability mechanisms (elections vs. exits).

a. Institutional purposes: politics vs. profits

The . . . most common concern is that public and private institutions pursue different purposes: politics and profits. . . . [G]overnments pursue political and legal goals more than economic gains; . . . corporations [do the reverse]. Almost every policy argument against the privatization of . . . force begins by carving out this basic boundary . . . . [C]ritics often call public [employees] “civil servants,” “bureaucrats,” and “professionals,” and private workers[] “employees,” “contractors” and “mercenaries.” . . . [T]hey draw upon a long line of liberal thinkers . . . who have articulated similar distinctions between public and private institutions and spheres.

In the simplest accountability arguments . . . there is little else than these two points. From such modest beginnings, critics leap to a verdict: Since the goals of punishment, policing, and military force are political and legal, the goals of punishment, policing, and military organizations must be political and legal too. Politics and profits must not be mixed. [Otherwise, c]orporations [will] trade public aspirations for private rewards.
b. Accountability mechanisms: elections vs. exits

The second concern is . . . more sophisticated . . . : States and markets rely . . . upon different accountability mechanisms. . . . [G]overnments react to the threat of “elections” and corporations react to the threat of “exits” [i.e., shareholders selling shares or customers stopping purchases]. . . . [E]lectoral accountability is public, but market accountability is not. In an election, each citizen can vote . . . once; in a market, only customers and shareholders can exit, and each exit is counted in dollars and cents. In short, public accountability is shared, whereas market accountability is sold.

2. Responses

   a. Accountability tradeoff

When public accountability objections are stated in such simple terms, responses are easy to find. For our purposes, let us focus on one that is already familiar by now: Notice how these two arguments stumble when they are applied to situations in which the supply of force is private and the demand for force is public.

The first argument sets forth a distinction between the purposes of governments and corporations. But . . . [c]orporations do not [just] pursue profits; they pursue profits by creating the kind of supply that satisfies demand. Just as governments pursue votes by pleasing potential voters, corporations pursue profits by pleasing potential customers and shareholders. They deliver goods and services of desirable quantities and qualities, at desirable prices . . . . If we want to compare apples to apples, we should say that governments pursue the goals of voters, whereas corporations pursue the goals of customers and shareholders. But this is where the second, more sophisticated argument kicks in: Accountability bought by customers and shareholders is not the same as accountability shared by voters . . . .

This “mechanisms” argument . . . is an important argument that operates within important limits. . . . I want to emphasize two: First, the mechanisms argument obscures an important tradeoff between accountability and publicness. When we compare elections and exits, we should weigh the pros and cons on both sides. It is true, as the mechanisms arguments suggest, that elections are more public than exits. But it is also true that elections are less accountable than exits.

By itself, this point probably seems trivial. The thrust of the public accountability argument is publicness, not accountability. This point, however, does not exist by itself. It stands next to another, more important point, to which I turn now. Just like the last argument, this one applies weakly to cases of public demand.

   b. Public demand

The two public accountability arguments—the purposes and mechanisms objections—suggest that the choice between public and private is a choice between politics and profits, elections and exits, voters and customers. But these distinctions start to founder before an already familiar foe—the public demand for force. Recall that governments are
the primary customers—in many cases, the only customers—of private punishment, policing, and military corporations. When governments are customers, the choices between customers and voters, elections and exits, and politics and profits are no longer so stark. In these contexts, the purposes and mechanisms arguments become rather weak.

As before, the clearest example is private prisons. . . . Voters elect politicians, who hire administrators, who hire private prison corporations. Administrators finance and allocate imprisonment. They act like customers: They write contracts, cut checks, and conduct inspections. If voters do not like the performance of public officials, they can . . . vot[e] for someone else. . . . [I]f public officials do not like the performance of private prisons, they can hold them accountable at the next negotiation, payment, or inspection by firing them and hiring someone else. Rather than swapping politics for profits, or elections for exits, the model of private supply and public demand marries the two. It pursues political and legal goals through economical means.

This is the customary response from privatization advocates. This give-and-take, however, is hardly the end of the matter. In some respects, the public demand response is equally naïve. It assumes the model of the perfect market, ignoring the possibility that private punishment, policing, and military markets might “fail.”

Consider these three points: First, the public demand model presumes that governments are virtuous and diligent customers, who are willing and able to protect the human rights of inmates, suspects, and combatants. Second, the public demand model glosses over the distinction between governments and shareholders. It presumes a perfect barrier between the government’s control of demand and the shareholder’s control of supply, so that the latter is not influenced, manipulated, or corrupted by the former. Third, it presumes that the marriage of private supply and public demand does not spawn other market failures, such as inadequate competition and agency costs. Whether these constitute fundamental objections to the privatization of force remains to be seen. Clearly, though, simple distinctions between “politics” and “profits” or “elections” and “exits” are inadequate.

E. Human Rights

1. Argument

The fourth argument against the privatization of force concerns the dangers of excessive and arbitrary uses of force. The central claim is that private punishment, policing, and military corporations violate human rights more often than [their public counterparts]. . . . [Most such arguments rely primarily on the logic of other arguments collected here.] But . . . one strain of human rights talk . . . is independent: anti-majoritarian arguments.

In anti-majoritarian arguments, critics attack the public demand model more directly. This critique is most often applied to the private prison market, in which governments are the only consumers . . . . [C]ritics claim that when governments hire private prisons, they are not interested in safeguarding the rights of inmates . . . . Instead, they are interested in cutting costs. For these critics, “privatization” . . . signifies the state’s subordination of inmates’ rights to . . . temporary political and economic victories. . . .
Anti-majoritarian arguments typically begin with historical narratives[,] . . . draw[ing] . . . analogies between nineteenth-century convict leasing and twentieth-century private prisons[,] and] provid[ing] left-wing histories of the development of the private prison industry . . . . When they list the penal reforms of the 1970s and 1980s, they often note that these reforms enjoyed widespread popular support [but] strongly imply that these reforms were unjustified . . . and motivated by . . . racism and greed. . . . This . . . talk challenges the premise of prison privatization—that governments need to construct, manage, and operate more prisons quicker and thus, need to turn to the private sector.

. . . [T]hese . . . narratives . . . highlight the way governments allow . . . politics to trump human rights. They draw a line between the interests and influence of voters and victims. They testify to the bad faith, or the weak will, of public officials. They suggest that the principles of privatization are democracy and downsizing rather than the protection of inmate rights. . . . [They] question an important assumption of the public demand model of privatization— . . . that public officials are virtuous, diligent consumers of force.

. . . [H]uman rights arguments . . . emphasize two types of dangers: First, companies may intentionally mistreat inmates . . . to cut costs, and second, companies may cut back on labor expenses, leaving inmates . . . at the mercy of unscreened, untrained, understaffed, and underpaid employees. . . . [T]he parade of pitfalls is horrific . . . .

2. Response

There are three major stumbling blocks for the anti-majoritarian strain of human rights arguments—evidence, style, and reform. . . . First, [the claim] that private punishment, policing, and military corporations violate human rights more often than public punishment, policing, and military institutions . . . has been vigorously and carefully contested, and there is very little empirical evidence to support it. Second, there is a stylistic tension . . . : They criticize governments harshly while urging for more public controls. They criticize the human rights policies of public legislative, judicial and administrative institutions; nevertheless, they defend the human rights practices of public punishment, policing, and military institutions. Of course, this is just a tension, not a contradiction. There could be good reasons to trust public prisons, police, and armies more than legislators, judges, and administrators. The problem is that [these critics] rarely offer such reasons, so the tension is rarely resolved. As a result, they often slide from arguments against private punishment, policing, and military corporations, into arguments against the exercise of punishment, policing, or military force itself.

This leads to the third obstacle . . . : If today’s privatization effort is motivated by the wrong objectives, why must we prohibit privatization instead of reforming it? Let us assume that . . . the anti-majoritarian critique is correct: When today’s governments privatize force, they sacrifice human rights to democracy and downsizing. . . . [W]e could assume that this majoritarian dynamic is a natural, necessary fact about governments and therefore prohibit governments from hiring private prisons, police, and armies at all. [Or], we could try to protect victims from voters. We could reform the privatization of force by reforming the demand-supply relationship between governments and corporations. We
could look for concrete ways to motivate public officials and private employees to respect and protect the rights of inmates, suspects, and combatants.

This road has not been traveled by human rights critics, but it has been positively trammeled by legal scholars. In law reviews, scholars have often observed that our most significant human rights safeguards have been designed to protect against abuses committed by governments rather than corporations. As a result, they have argued, the activities of private punishment, policing, and military corporations often fall through legislative, judicial, and administrative cracks.

Let us call these “loophole” arguments. In such arguments, legal scholars rarely suggest that loopholes are reasons to ban the privatization of punishment, policing, and military force. Far more often, they argue that we should regulate private prisons, police, and armies by closing the gaps . . . . They argue, for example, that (1) private prisons should not enjoy immunity from federal lawsuits, (2) private police searches should be governed by the Fourth Amendment and thus justified by probable cause, and (3) mobilized defense contractors and PMC soldiers should be classified as lawful “combatants” under the Geneva Convention and bound by international laws of war.

. . . [I]t is easy to [also] imagine . . . solutions that draw upon the powers of legislators and administrators as well. Contracts could be drafted to include more rights-conscious performance incentives, reporting and monitoring requirements, and insurance, damages, and termination options. Statutes and regulations could be drafted to require public agencies to monitor our private prisons, police, and armies more closely. Or they could empower human rights groups . . . to do so instead.

Would any of these reforms work? . . . Would regulated corporations exercise punishment, policing, and military force more or less callously than governments? . . . Although these questions are familiar and inevitable, anti-majoritarian arguments provide no useful tools to resolve them. When it comes to private supply, these arguments make a strong case for reform, but a weak case for prohibition. They offer no proof that the model of private supply and public demand is broken beyond repair.

Yet they do suggest that reforms would be hard won, if not unwinnable. . . . [I]n the real world, the privatization of punishment, policing, and military force is not likely to be accompanied by human rights reforms. In political terms, the privatization of force is about democracy and downsizing, not human rights. We should not expect governments to reform private punishment, policing, and military markets by themselves. . . . This is the greatest strength of anti-majoritarian arguments: They offer the historical, practical reminder that voters rarely protect victims, so that democratic governments are inconstant guardians of human rights.

This is a useful admonition. But it criticizes the privatization of force qua popular politics rather than privatization qua privatization. It fails to articulate any structural, systematic link between the privatization of force and the regulatory failures of governments.
F. Industrial Influence

1. Argument

a. Influence

This link is the subject of the “industrial influence” argument, the fifth argument against the privatization of force. Like the last one, this criticism is focused primarily on the private prison and defense contracting industries, which rely exclusively on the model of private supply and public demand. In principle, however, it is easily adaptable to the public demand sector of the private policing and private military industries. The industrial influence argument focuses on the powerful interest groups formed by private punishment, policing, and military industries. Critics argue that these interest groups spend money to influence, capture, and corrupt our legislative, judicial, and administrative institutions. Corporations invest in lobbying, donations, and bribes. They persuade public officials to purchase services from private punishment, policing, and military corporations, and to deregulate private punishment, policing, and military markets. They sponsor pro-industry studies, seminars, and conferences, exchange public and private personnel, and finance campaigns. They cultivate contacts, friendships, and professional debts. In other words, they barter influence.

To economists, this is yet another example of market failure, which they call “rent-seeking” behavior: Corporations waste public and private funds, spending shareholder money to squeeze excessive benefits from public projects. For most critics, industrial influence is a problem of deliberation, and thus, a problem of democracy. [They] are concerned that privatization leads politicians to think about politics in selfish and irrational ways. First, after enough donations and bribes, officials might think more of themselves and less of the public at large. They might support punishment, policing, and military policies that advance their own careers or line their own pockets. Second, after enough lobbying, officials might be more easily persuaded by sham evidence and shoddy arguments.

The most harrowing and persuasive industrial influence arguments look to the past and the future. President Eisenhower’s warning about the military-industrial complex is echoed in talk of the “prison-industrial complex.” In the future, critics predict that these industries will buy more and more political influence and become a primary shaper of the popular demand for punishment, policing, and military force. They will fund politicians and administrators who are tough on crime and tougher on terrorism. To feed market growth and produce investment returns, corporations will fund research and reform efforts that target newer, larger “demographics” of inmates, suspects, and foes. Payments and allocations will be motivated by self-interest and fear.

b. Comparison

Like every argument against privatization, this one is comparative; the problem of influence is not unique to markets. Understandably, wardens, police chiefs, and generals...
believe that they provide vital services to the public. . . . [T]hey[, too,] lobby politicians and administrators for bigger budgets and broader mandates.

It is easy to imagine [how] such partnerships between public supply and public demand go astray in everyday government. . . . [W]e must admit that some wardens, police commissioners, and generals are likely to trade political favors with legislatures and regulators in pursuit of professional and personal goals.

The challenge is to explain why the problem of influence would be worse in “private-public” partnerships than in “public-public” partnerships—why private prisons, police, and armies would extract more rents from legislatures than [their public counterparts]. More often than not, industrial influence arguments do not bother to prove this critical point. But the logic of such arguments suggests a plausible answer: Whereas public institutions are bound to obey public budgets, corporations may reinvest profits into the political process itself. In particular, corporations have the capacity to exchange revenues with legislatures by trading political contributions for government contracts. When corporations and governments join forces, a distinctive type of tender—political contributions—greases the skids for influence peddlers.

. . . [B]etween public institutions, the buying and selling of influence is a strictly black market affair. Wardens, police chiefs, and generals . . . must spend institutional funds on prisons, police, and armies . . . . As public officials, they can pull strings but they cannot cut checks. As a result, they must barter for political influence instead of buying it.

In partnerships between corporations and governments, by contrast, there is a relatively open market for political power. In the United States, for example, corporations may establish, administer, and solicit contributions to “political action committees” . . . . For private punishment, policing, and military corporations, such political action committees are rational and legal investment vehicles. . . .

Of course, such committees can be defended as organized efforts to express political interests and invigorate political debates. But when we compare public and private institutions, we must prepare for the worst of both worlds. . . . We may no longer imagine private contributions to ideological allies; we must contemplate winks and nods between politicians and entrepreneurs. Politicians, administrators, and contractors may exchange tit-for-tat, payment-for-payment, contribution-for-contract. In return for “reasonable” donations, corporations may tap into taxpayers’ pockets.

2. Response

[This argument] seems fabulous. It achieves what the others could not: It explains one of the structural, systematic ways in which the privatization of punishment, policing, and military force engenders regulatory failures. More than the earlier arguments, it confronts the public demand model head on. It suggests that governments make better managers of markets than they make customers in markets. More specifically, it belies the second feature of the public demand model—the perfect barrier between the shareholder’s control of supply and the government’s control of demand. It demonstrates that supply
does not simply “satisfy” demand. Shareholders, managers, and employees pursue their own interests, instead of passively satisfying consumers. When markets expand, corporations develop the ability to influence, manipulate, and corrupt public demand. The more that we grant them control of provision and exercise, the more that they purchase control of payment and allocation. Slowly, private supply warps public demand.

This is serious business in the context of punishment, policing, and military markets[, which involve] . . . elemental decisions of equality, freedom, and justice. . . .

Many . . . critics claim that these problems are already upon us. They claim that our rising prison populations are driven, in part, by the private prison industry and that our rising security costs are driven, in part, by private security firms. . . . But the strongest case for the industrial influence argument lies far in the future, in a world in which the privatization of force is much more advanced. If we had enormous, powerful punishment, policing, and military industries, then we would really have something to worry about.

But does the industrial influence argument really require [the prohibition of privatization]? In theoretical terms it does not. Like the human rights argument, it is not an argument against the privatization of punishment, policing, and military force by itself. It is an argument against the combination of corporate campaign contributions and the privatization of force. If we curb the problem of corporate campaign finance, then we curb the problem of industrial influence. In other words, we face [a] practical fork in the road. Which policy could be changed easier, quicker, and more completely? Current events offer only a crumb of evidence on both sides: In the past three years, Congress has passed the McCain-Feingold Act and the Aviation and Transportation Security Act. The former victory was hard won; the latter was easily won, yet motivated by the exceptional events of [9/11]. [To tackle the questions] which was the greater exception and which was the more plausible rule, let alone which one is more likely to “work[,]” . . . would lead us far afield. The question remains: Is there anything fundamentally wrong with the privatization of punishment, policing, and military force per se?

G. Market Failure

1. Arguments

Critics take up this question more directly in a sixth set of arguments. They argue that the public demand model produces two final forms of market failure—inadequate competition and agency costs.

a. Inadequate competition

The inadequate competition argument is closely related to the industrial influence argument. . . . Critics argue that [with private prisons and defense contractors], only a few large corporations dominate the field. In economics, this is called “oligopoly”—a market in which a small group of suppliers excludes entrants and raises prices.

. . . [C]ritics argue that the root cause of oligopoly is the concept of a “natural monopoly”: a market in which high “fixed” costs and “sunk” costs diminish incentives to
enter or exit. For our purposes, fixed and sunk costs involve supply-side investments in highly specialized construction, materiel, and perhaps expertise. Costs are “fixed” when they do not decline as provision expands; costs are “sunk” when they cannot be recovered upon exit.

Such costs give rise to inadequate competition in several ways. Most of all, they diminish the incentives for new firms to enter, old firms to exit, and customers to switch midstream. Only a few firms have the necessary resources and expertise to compete for large contracts. Contracts are long-term. Cancellations are rare, costly, and contested. Prices stabilize or steadily increase. Strong ties exist between public officials and private contractors, who depend on one another to survive and thrive. Rent-seeking is rampant.

b. Agency costs

The agency argument is more complex. It comes from the economic theory of agency relationships, in which principals hire agents to do things for them. Agency costs arise when agents do not do what principals desire. In agency arguments, the principals are citizens and/or governments, and the agents are public and private punishment, policing, and military corporations. As one commentator explains, there are several common causes of agency costs: “[T]he agent’s interests do not entirely coincide with those of the principal; the principal does not have complete control over the agent; the agent has only partial information about the principal’s interests; and the principal has only partial information about the agent’s behavior.”

Agency theory describes several different kinds of agency relationships which are each suited to minimize different kinds of agency costs. In this context, the two relevant kinds are outsourcing and employment relationships. In “outsourcing” relationships, principals hire agents to deliver results. In “employment” relationships, principals hire agents to obey orders. Outsourcing works well when principals have simple, predictable, certain goals, and do not care too much how agents achieve them. Agents know ex ante what principals want them to do, and principals know ex post what agents have done. In contrast, employment works well when principals have complex, unpredictable, uncertain goals, and care a lot about how agents achieve them. . . .

Agency critics observe that the goals of punishment, policing, and military organizations are exceedingly complex, unpredictable, and uncertain. Moreover, we care immensely how these goals are achieved. As a result, governments encounter problems when they attempt to outsource punishment, policing, and military force. Contracts are difficult to write, terms are hard to negotiate, and compliance is hard to monitor. By introducing the profit motive, we increase agency costs. We grant agents more room to run amok; we must work harder to stop them.

2. Responses

a. Economics

Finally we have two astute objections to the public demand model itself. Like the industrial influence argument, [they] show how private supply can shake free from public
demand. They suggest two different ways in which punishment, policing, and military corporations could frustrate, escape, or ignore the public demand for force. But these two arguments are better than the last one. They are not about defects . . . in our political and legal processes. They are about defects . . . in private punishment, policing, and military markets themselves. The inadequate competition argument attributes market failure to the nature of punishment, policing, and military tasks; the agency costs argument attributes market failure to the nature of punishment, policing, and military goals. The competition argument focuses on the costly nature of private supply; the agency argument focuses on the complex, unpredictable, and uncertain nature of public demand.

But if we mean to conduct a liberal analysis of the privatization of force, these criticisms provide funny hooks upon which to hang our hats. When we started this analysis, we conceded the economic efficiency argument. Then we moved beyond economic questions to explore a series of traditional liberal values—equality, security, democracy, and liberty. Among these values we found few fundamental concerns. Now, we finally find two inherent objections to the privatization of force, but we have apparently come full circle. These are economic arguments, unadorned by political and legal values. They argue that punishment, policing, and military markets minimize competition and maximize agency costs. Market failure arguments pull out the two pillars of the free market model: competition and information.

In doing so, they land us back where we started: in the heated empirical debate over whether corporations “really” provide punishment, policing, and military force more efficiently than governments. For critics of privatization, this is enemy terrain. . . .

I grant that these two market failure arguments have serious political and legal implications. When we are talking about punishment, policing, and military force, every economic argument does. Both arguments suggest structural ways in which private punishment, policing, and military corporations could betray our public demands and develop immunities to the possibilities of public negotiations, inspections, and sanctions. They remind us that when markets fail, the accountability of corporations fails too. If these arguments are correct, then private punishment, policing, and military corporations would not do the things that we want them to do. In this sense, these two market failures pose the widest political and legal threats. They threaten equality, security, democracy, and human rights, all at once.

In the end, however, these market failure arguments remain tethered to economic roots. . . . [E]ach argument is limited by the horizon of the economic perspective. The competition argument is too narrow-minded; it assumes that the demand for force is not malleable and, thus, it overlooks a simple path to reform. The agency argument is much better, but it is still too broad. It fails to explain the singular significance of punishment, policing, and military force. But these are cryptic claims that must be explained.

*b. Competition*

To consider the limits of the competition argument, let us focus on the problem of private prisons. Critics claim that fixed and sunk costs are unusually high in this industry which
discourages new firms from entering, old firms from exiting, and public officials from switching firms. The major fixed and sunk costs in the industry are the construction, design, and lease purchase of prisons.

... [S]ecuring public funding for prison construction is a lengthy and uncertain process. Voters often support long prison sentences, but they do not often support large prison bonds. Even when they do, public financing involves budget hearings and design competitions, which private financing avoids. So governments enter a “lease purchase” agreement. The private prison corporation raises the capital, designs the physical structure, and finances the construction project. The state leases the prison from the corporation until the contract expires.

Is any of this necessary? Only if you assume that demand is fixed. It is easy to imagine a world in which lease purchase agreements are illegal . . . [and] . . . officials [must] secure public funding. There are two reasons to prefer this world: First, some might argue that this system would be more just. Voters would be forced to reconcile allocations and payments. If they refused to pay for long sentences, they would [have] to abandon them.

But let us put that point aside, and turn to the second reason . . . . [A] publicly financed, privately supplied prison industry [might] be more competitive than the one we have now. Firms would not bear the burdens of financing construction, design, and ownership. Both corporations and governments could enter and exit contracts more easily. Would private prisons still be cheaper? [Perhaps]: First, lease-purchase agreements often cost more than public bonds because corporations are less effective risk spreaders than governments. Second, construction, design, and ownership are not the only costs of imprisonment, nor even the most significant ones. Labor represents roughly sixty percent of the overall cost of imprisonment, and labor is neither a “fixed” nor a “sunk” cost.

Cutting labor costs raises other problems, of course, which we have already seen. If labor costs drop too low, people are placed at risk: inmates, guards, and, if escapes increase, everyone else. But, as we have already seen, that argument does not prove much. This problem can also be resolved through public demand reforms. It proves that governments should pay for (and allocate) high quality labor, but it does not prove that governments should provide labor at all. Taken together, the two arguments prove only that any form of imprisonment—public, private, or mixed—imposes significant costs. . . .

c. Agency

Since the agency argument is more complex, the economic limitations of this argument are harder to describe. It is a remarkably ambitious argument. It purports to identify those criteria upon which any kind of organization should choose between employment and outsourcing. . . . [I]t offers a complete theory of our public choices between public and private supply[,]. . . . incorporat[ing] the complexity, predictability, and certainty of ends, and the relative importance of ends versus means.

It is the most thoughtful argument yet. . . . No reform proposals could hope to mitigate the inherent complexity, unpredictability, and uncertainty of our punishment, policing,
and military policies. No one could convince us that in these three endeavors, methods are not important.

But I do not think that this argument conveys everything that we want to say about the supply of punishment, policing, and military force. It is too generic. Consider the first three criteria: complexity, predictability, and certainty. Imagine that public officials decide to fund a few projects [that] are complex, unpredictable, and uncertain. The argument recommends that they establish a public agency of “civil servants,” “bureaucrats,” or “professionals” to pursue these goals.

So which are our most complex, unpredictable, and uncertain public endeavors? High on anyone’s list would have to be publicly funded programs in scientific research, humanities research, and the fine arts. The argument suggests that the supply of these programs should be public. Is it? Some supply is public, some is private, and some is non-profit, quasi-profit, and various other middling shades. The goals of research and arts programs are at least as complex, unpredictable, and uncertain as the goals of punishment, policing, and military institutions. These criteria suggest that the case for the public monopolization of ownership, management, and employment should be at least as strong in the fields of research, art, and force.

I think not. I think that there is some wisdom to the idea that the state has, must have, or should have a “monopoly of force.” I do not think that there is much wisdom in the idea that the state has, must have, or should have a monopoly of research or art.

But these counterexamples are merely illustrative. My point is conceptual: I do not think that the state must monopolize the pursuit of all complex, unpredictable, or uncertain goals in similar ways, to similar degrees. But I do think that the state must monopolize the supply of force. The question remains: Why force?

The final criterion of the agency argument hits the mark: the relative importance of ends versus means. This last criterion suggests an obvious and powerful distinction between force, on the one hand, and things like research and art, on the other hand. Only the most illiberal sensibilities could suggest that scholarly and artistic methods are more important than—or even equally important to—punishment, policing, and military methods. Liberals might believe that the goals of research, art, and force are all equally important, but they could hardly suggest that the specific means are as well.

We monopolize punishment, policing, and military force because, in these projects, methods and results are both profoundly important to us. We care not only that force is exercised, but also when, where, how, how much, and against whom. When we privatize punishment, policing, and military force, we surrender control of punishment, policing, and military methods. We surrender control of the supply-side decisions about who may use force as well as some of the decisions about when, where, how, how much, and against whom forced may be used. Although these are less critical than demand-side decisions, they are nonetheless decisions of great consequence and importance.
In the great tradition of utilitarian economics, this last argument is both structurally brilliant and morally vacant. It is structurally brilliant because it is true: We monopolize the exercise of force because it matters how force is exercised. It is morally vacant because it fails to ask why. Why is the exercise of force—instead of research, art, or anything else—so important? More specifically, why is the exercise of force—in addition to the allocation of force—so important? We are on the right track, but we have not yet cracked our chestnut: “The state must monopolize force.” What we lack is a liberal account of the special significance of force.

**H. Cultural Commodification**

*1. Argument*

Enter the seventh and final policy argument against the privatization of force. Like the others, [it] begins with the premise that governments and corporations pursue different purposes. Unlike the others, however, it does not make any predictions about how these different purposes cause different actions or create different results. The logic . . . is more cultural than instrumental[: When we privatize the exercise of punishment, policing, and military force, we privatize the social meaning of punishment, policing, and military force. On this view, the privatization of force engenders the commodification of force. . . .

The term “social meaning” describes an interpretation shared by a society; the term “commodification” describes the way that a society comes to interpret something as a commodity—as an object of economic exchange—and to imagine or recognize the existence of a market. . . .

The commodification argument hinges on two presumptions. The first is old hat: . . . governments pursue public purposes; corporations pursue private purposes. The second is new hat: . . . institutions and meanings are necessarily linked, such that when we change the purposes of institutions, we change the social meaning of institutional practices too. . . . When public officials exercise force, they convey public messages, such as “just say no,” “respect the rule of law,” and “don’t tread on us.” When private employees exercise force, they convey private messages, such as “the customer is always right” or “Force, Inc.” When private employees catch and punish criminals, defend borders, and wage wars, they “commodify” force. Society reads these actions as “about” the bottom line rather than “about” some political or legal ideal. Thus, the privatization of force “degrades” or “cheapens” exercises of force. It strips them of public significance.

Like human rights arguments, commodification arguments . . . focus on . . . victims . . .: inmates, suspects, and the casualties of war. First, . . . privatization sends the message that these victims are not human beings but commodities. Second, . . . privatization places cultural “distance” between the victims of force and the citizens of the liberal state. It signals our indifference to these victims, our refusal to take responsibility for [their] fates . . . . By adding an extra institutional layer between ourselves and our exercises of . . . force, we imply that these are no longer our arrests, our punishments, our wars, and thus no longer our suspects, our prisoners, our casualties. . . . [W]e wash our hands of our most violent deeds.
2. Responses

a. Conclusions

... Many ... have responded, “So what?” ... [C]ritics rarely bother to explain the normative importance of commodification. This is a shame, because there are several powerful ways to respond. The best answer suggests that exercises of force [should] convey important public messages, and that privatization scrambles the reception of these messages. Social meanings are not merely side effects of exercises of force; they are also one ... reason[] ... we exercise force. They represent one aspect of our ... policies, purposes, or goals. We exercise force not only to stop crimes and fight wars, but also to send public messages, create public symbols, and express public norms.

In liberal states ... force is exercised to send liberal messages and maintain liberal values. Liberal societies police and punish criminals to proclaim that crime is wrong and the social contract is right. They fight wars to affirm the ideals of the rule of law, democracy, and human rights. These social meanings are vital social resources; they are a large part of what binds us together and makes us behave. When we exercise ... force, we strengthen the cultural foundations of our [society].

Not always, to be sure. Just as we do not always catch suspects and win battles, we do not always get our messages across. But if the commodification argument is correct, then privatization will cause our ... messages to misfire more often. Even when we achieve our instrumental goals, people will attribute the “wrong” purposes to exercises of force, and our cultural goals will not be achieved. [Our] cultural fabric ... will weaken.

This is a definite improvement over the agency argument. It explains why the means of force, as opposed to the means of other collective endeavors, are especially important to us. Exercises of ... force ... convey[] vital messages. ... [T]hey both protect and destroy our most valued social goods: property, liberty, and life. They jeopardize and sacrifice in the name of ideals. These are intensely meaningful practices. ... 

This argument makes terrific sense. It suggests the most probable and plausible justification for our recent “federalization” of the airport security supply—to pronounce ... our national commitment to “homeland defense.” It helps us understand why we “read” exercises of punishment, policing, and military force—because they are designed to be read. Most importantly, the argument provides a purely non-economic, political objection to the model of private supply and public demand.

b. Cultural change

But the argument has a few fundamental flaws. We will start with the most straightforward point: When we talk about the social meaning of an exercise of “force,” we must consider [both] supply and demand. When we interpret an exercise of force, to determine whether it is “public” or “private,” we ask two kinds of questions. On the supply-side, we ask who provided and exercised it—who carried it out at the search, arrest, incarceration, or attack. On the demand-side, we ask who paid for and allocated it—who signed the contract, passed the sentence, issued the warrant, or declared the war.
The commodification argument suggests that when we change who exercises force, we change the purposes of force and, thus, the social meaning of force. Already, we see that the argument had softened some: In truth, publicly-allocated, privately-exercised force is not completely commodified. It is a muddle of public and private purposes and, thus, a mixture of public and private meanings. It is a mixed message.

But the argument still seems to stand: Mixed messages do not often blend into strong cultural cement. We read exercises of force . . . , in part, by asking who exercises force. We expect to find out that public officials exercise force. Thus, we expect to read public meanings from public exercises of force. When private employees exercise force, our expectations are frustrated. We read public/private meanings from private exercises. Vital messages are not conveyed. . . . The cultural pillars of the liberal state start to shake.

Or do they? For some reason, the commodification argument assumes that when institutions change, only the content of our interpretations change. [But w]hy would the methods or sources of our interpretations not change, too? . . . Suppose that for fifty years, public officials ignore the critics . . . [and] continue to hire private punishment, policing, and military corporations. . . . After fifty years, our entire prison population is housed in private prisons. In this world, what is the social meaning of imprisonment?

The commodification argument only sees one possibility, but I see two. In the first scenario, . . . [i]mprisonment is . . . commodified. In this world, when we interpret imprisonment, we ask ourselves both, “Who exercises imprisonment?” and “Who allocates imprisonment?” . . . [T]he social meaning of imprisonment is mixed . . . .

But in the second scenario, . . . [i]mprisonment is . . . not commodified. . . . [W]e have created new practices of cultural interpretation to suit our new practices of punishment. (Think of it as a cultural “reform,” although it would probably be neither intentional nor self-conscious.) When we interpret imprisonment in this world, we do not think at all about who exercises imprisonment. We just ask ourselves, “Who allocates imprisonment?” . . . We read public values from public allocations of imprisonment. . . . [W]e see the same old symbols: “Crime is wrong and the social contract is right.” We read nothing from private exercises of imprisonment. We recognize that private prisons have private purposes, but we do not consider this to be a salient cultural fact.

Could that happen? . . . [W]hen it comes to the task of cultural self-justification, . . . we are a remarkably adaptable species. We often perform backflips and cartwheels to “read” ourselves right. . . . I suspect that in some circles, this cultural shift is already well underway. I venture that there are already some legislators, judges, administrators, and entrepreneurs who actually and honestly do not believe that “private” imprisonment is significantly different from “public” imprisonment in cultural terms. When they consider the social meaning of imprisonment, they think about the public character of legislators, prosecutors, judges, and juries rather than the private character of wardens and guards.

But my argument does not hinge on these empirical claims. For my response, it is enough to say that this could happen. . . . Would this culture be “liberal”? Commodification critics want to say “no.” They purport to identify necessary yet mysterious links between
public purposes, public meanings, and public responsibilities: Since private prison guards have private purposes, these purposes must have social significance. It must matter who exercises force[, or else] this must be a callous society in which no one cares about the exercise of imprisonment. In this society, private prisons deprive inmates, but the public remains indifferent. We tell ourselves that they are not our prisons and, thus, not our prisoners. But . . . [w]here are such imperatives written? The Liberal Book of Culture?

Liberal societies do not always and everywhere interpret “privatization” as culturally significant. Sometimes, private purposes exist within a larger institutional framework of public purposes and public responsibility. . . . Our [“public”] guards, police, and soldiers are not drafted, and they do not work for free. They are a volunteer corps that applies to work and gets paid for it. Presumably, these payments introduce some private purposes into our public prisons, police departments, and armies. [But] we dismiss these purposes as culturally irrelevant. Why? Because we believe that “public” guards, police, and soldiers work within a larger institutional framework of public ownership, oversight, bureaucracy, and professionalism. In cultural terms, could we not say something similar of private prisons, police, and armies themselves? Could we not say that they are contained within a larger system of public executive, legislative, and judicial institutions, which have public purposes, and impose public responsibilities? Could we not deem our payments to private prisons, police, and armies to be culturally insignificant?

There are good reasons to say no, to say that public demand cannot control private supply, but we have already seen them. They are economic and political reasons, not cultural ones. When we reject this possibility on cultural terms, we must offer cultural evidence or cultural arguments. If we do not, we practice cultural imperialism. We presume that public ownership, management, and employment “really” mean one thing and private ownership, management, and employment “really” mean something else. . . .

In commodification arguments, there is a fair amount of this conclusory talk. . . . [T]he privatization of force is an “unseemly,” “inappropriate” delegation of an “inherently public function,” a function that is “the quintessence of the state’s sovereign power.” However, these statements appear as rabbits from hats. How do we know what functions are “inherently public”? What delegations are “unseemly” and “inappropriate”? What privatization “means”? . . . [T]hese . . . [are] circular justifications for cultural critiques.

I. Limits and Flaws

. . . [T]hese questions are tough love. . . . I share the vague intuition that private punishment, policing, and military corporations are unseemly, perhaps even illiberal. . . . Today, the monopoly thesis is firmly rooted in our political culture, widely regarded as a building block of our public values. . . . [But] we should no longer be satisfied with pat restatements of ancient ideas. . . . If we mean to advance our public policy debates . . ., we must re-examine and re-justify the old idea that the state must monopolize force. More specifically, we must distinguish between monopolies of supply and demand.

By now, it should be clear that this is . . . what our policy debates have failed to do. [Each of the critiques—]economic efficiency, egalitarian distribution, public goods, public
accountability, human rights, industrial influence, market failure, and cultural commodification—. . . voices . . . important concerns [and] provides strong . . . reasons to be skeptical of the private demand for punishment, policing, and military force.

Yet each one applies more weakly and unclearly to the private supply of force. Once we assume that government will more or less finance and allocate force, it becomes much less clear whether, how, and how well these arguments apply. Perhaps governments should actually provide and exercise force too. Perhaps corporations should provide and exercise force instead, under the careful watch of public institutions, bolstered by a commonplace package of legal, economic, and redistributive reforms. This is the model of private supply and public demand. In our public policy debates, when demand is public and supply is private, the devil slips into the empirical details.

We should ask ourselves how we got into this analytical mess . . . . I have identified three basic flaws in our public policy debates, which limit the depth of our current analyses: Our arguments are (1) too boilerplate, (2) too indiscriminate, and (3) too “stuck” in our present situation, too mired in the modern administrative state.

1. Boilerplate Arguments

First, our arguments are boilerplate. They look surprisingly like the public/private debates over non-coercive endeavors, such as education, health care, housing, or even water. It is as if we have taken our old template for “public/private” debates and imposed it upon ourselves yet again. We have turned what might have been a distinctive controversy into a classic clash between utilitarians and liberal-democrats. . . .

2. Indiscriminate Arguments

Perhaps as a result, our arguments are indiscriminate. Each one gestures or aspires toward a theory of everything . . . . While privatization advocates assure us, “it is all about efficiency,” critics argue that it is “really” all about “democracy,” “rights,” “accountability,” or “culture.” They are all doubly wrong: In the practical sense, it is about all of these things. In the philosophical sense, it is not about any of these things.

. . . None of these principles offer any clear or compelling ways to differentiate the private supply of punishment, policing, and military force from one another. In our policy debates, “force” often functions as no more than an extreme example of a very general idea of what government is supposed to do. In this way, the debate has lost touch with the instinctual, visceral energy that we sense in that vague liberal verdict: “The state must monopolize force.” . . .

3. Modern Administrative Arguments

. . . It is too easy to mock the state-of-nature story and the monopoly thesis as idle fictions. . . . Some fictions—even apparently naïve or nonsensical ones—reveal great truths. . . . [I]n this context[, t]hey help us reveal the third flaw of our public policy debates: They are too deeply attached to present realities. In the grand scheme of things, our present realities are somewhat mundane. Private prisons house less than ten percent
of our adult prison population. Private police are everywhere, but they are not running our local police departments, let alone the FBI or the CIA. Private armies are fighting small conflicts, but in the biggest battles, they fight alongside public armies or they do not fight at all. The change has been striking, but as we have seen, it has hardly begun.

In this atmosphere, critics are forced to fight at the margins. They must make arguments that apply to the somewhat modest private punishment, policing, and military markets of today. They must object to the kinds of private prisons, police, and armies that exist today and the kinds that may emerge tomorrow. They make these arguments in the shadow of an imperious institution—the modern, administrative, liberal state. At every turn, they must confront the possibility of public, liberal reforms: redistributions, contracts, inspections, and regulations. They are forced to make empirical claims about which institutions are most likely to fail and which reforms are most likely to succeed.

V. THE LIBERAL THEORY OF FORCE

The state of nature story helps us escape these earthly realities. It shakes us free from our worldly assumptions, pushing us beyond our normative horizons. . . . When we imagine a condition of anarchy, we push the privatization principle all the way, to its logical end. . . . [T]he idea of a . . . state, in which the demand for punishment, policing, and military force is completely public, but the supply of punishment, policing, and military force is completely private[,] . . . is our first thought experiment—the “ultraminimal state.” . . .

The argument proceeds in four stages: First, we develop the thought experiment of the ultraminimal state . . . . Second, we trace the first step of the liberal state’s evolution from the “ultraminimal state” to the “military state,” a close cousin in almost every respect. Third, we generalize our first argument into a few basic principles, which we then use to pinpoint how far the liberal state lies beyond the military state. This wraps up the problem of private supply. Lastly, we rethink the problem of private demand. . . .

C. The Military State

1. The Problem of Loyalty: Take One

a. The ultraminimal state

. . . Imagine that after decades of downsizing and privatizing government, Americans have eliminated everything but our most essential administrative agencies, courts, legislatures, and executive offices. We still rely on government for the same array of social services, but we outsource them all. We pay private firms to do public works in exchange for taxpayer funds.

As a result, our massive, modern administrative state has been transformed into a lean, well oiled machine. It is no longer just the post office that turns a profit. In all public projects, we get more bang for our collective bucks. Thanks to a healthy dose of liberal reforms—redistributions, contracts, inspections, and regulations—we receive these services where, when, and how we want them.
But this is an *ultra*-minimal state, so it . . . is also without three hallmarks of the modern state: public punishment, policing, and military institutions. Private military corporations defend boundaries, protect allies, and fight wars. Private police firms find and catch suspects. Private prison companies detain and execute criminals. . . . [T]he supply of force is completely private.

. . . This is not quite anarchy. . . . It seems like a state. The state *demands* force; the market *supplies* it. The state uses political, legal, and economic mechanisms to keep private punishment, policing, and military corporations in line. . . .

. . . Initially the plot unfolds as planned. Boundaries are defended; allies are protected; wars are fought. Suspects are found and caught. Criminals are imprisoned, and if you prefer to imagine it, the worst criminals are executed. The system runs smoothly: The state pays and the market provides.

One day, however, something snaps. Somewhere in the upper echelons of the private military industry, . . . someone perceives an illicit opportunity and hatches a criminal scheme [that] subvert[s] the contracts, inspections, regulations, and redistributions of the ultraminimal state, and . . . generate[s] income or the prospect of income to come. It could be a simple theft[,] a small conspiracy[,] . . . a military intervention in a disputed election . . . [, or a] civil war . . . . It could involve an endless list of potential partners: criminals, terrorists, rebels, enemy states, and other corporations.

Let us call this the problem of loyalty. If the crisis remains small, the ultraminimal state could suppress it. It could pay private police to . . . apprehend . . . the conspirators [o]r . . . pay loyal private armies to quash the rebellion and defend the state. Suppose, however, that the crisis spreads, like a virus, through the private military industry. Either by collusion, acquisition, merger, competition, or violent conflict, several of the most powerful, profitable private military corporations come together . . . [and] form a private monopoly of the supply of military force within the territory of the ultraminimal state. . . . [T]hey form a dominant military organization, which we will call an “MO.” Within the state’s borders, the MO is *dominant*: It is the organization most capable of using acts and threats of physical coercion to subordinate individuals and institutions.

Now . . . [t]he MO . . . [could] violent[ly] overthrow . . . the ultraminimal state and . . . establish[] “State, Inc.” Of course, the ultraminimal state would be crushed instantly by the MO. If politicians, judges, and bureaucrats refused to pay the MO for services, tried to “fire” the MO, or imposed financial or criminal sanctions upon the MO, they would be forced off stage. The state’s impressive array of political, legal, and economic reforms would be worth little more than the pages on which they were written. In short order, the MO would install a new regime. Thus ends the story of the ultraminimal state.

*b. The military state*

. . . Now, [let us] imagine that individuals created a “military state” rather than an ultraminimal state. In the military state, the supply of military force is public: Military

* [For convenience, I have moved these last few paragraphs into section C from section B.1. —Ed.]
force is publicly provided and exercised as well as publicly financed and allocated. In every other way, the military state is just like the ultraminimal state: Although the military state owns, manages, and staffs military institutions, it still contracts out policing and punishment to the market.

We must be clear on one point: Our “military state” is not a traditional “military state,” in the sense [of being totalitarian or authoritarian or ruled by a military junta]. It is a state founded by rational, self-interested individuals. It is constructed in an effort to . . . pursue the liberal ideals of democracy, liberty, and justice. It is precisely like the ultraminimal state, with one important exception: It monopolizes the supply of military force.

. . . [I]t appears that this distinction makes all the difference: The military state seems to solve the problem of loyalty. If any other entity, such as an angry mob, a private policing corporation, or another state, were to try anything funny . . . [, t]he military state could crush any treasonous schemes. It could rely on its monopoly of military force to stifle any threats of revolution, civil war, or conquest. Unlike the ultraminimal state, it could establish and maintain order.

2. The Problem of Loyalty: Take Two

Perhaps you already sense the truism: The obvious answer treats “the state” and “the market” as real institutions, rather than metaphors, against my earlier advice. Without argument, it presumes that public military institutions are loyal—or at least, more loyal than private military corporations. . . . After all, “the state” is “the state,” and it is neither rational nor self-interested to betray oneself. In the “military state,” “the military” becomes part of “the state.” . . . In this answer, the concept of the “military state” literally dissolves the problem of loyalty.

. . . [T]his kind of wordplay is conclusory. . . . [Instead, w]e can reconsider the ultraminimal state and the military state in terms of two different relationships between supply and demand organizations:

**Figure 4**

<table>
<thead>
<tr>
<th>Demand</th>
<th>Ultraminimal state</th>
<th>Military state</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative, judicial, legislative, and executive institutions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supply</td>
<td>Private military corporations</td>
<td>Public military institutions</td>
</tr>
</tbody>
</table>

Draft — do not distribute!
This more involved formulation of the problem of loyalty has several significant advantages. First, it reconceptualizes “the state” and “the market” as institutional relationships, rather than actual institutions. From this more accurate perspective, it is easy to see through the sophistry of the obvious answer. We can no longer say, for example, that the “military state” possesses a “monopoly of military force,” or that the “ultraminimal state” lacks a “monopoly of military force.” We must speak more specifically of organizations, monopolies, and the relationships between them.

Second, this formulation reveals that the relevant players are roughly the same in the two states. In each case, there is a network of administrative, judicial, legislative, and executive institutions, which monopolizes the demand for military force. In other words, there is a deliberative government organization, or a “GO.” In each case, there is a network of military institutions . . . which monopolizes the supply of military force. So there is an MO in both states. In the ultraminimal state, the MO is private; in the military state, the MO is public. We are no longer looking for loyalists and traitors. We are looking for more or less durable relationships between the MO’s monopoly of supply and the GO’s monopoly of demand.

Third, this formulation suggests that the problem of loyalty is no paper tiger, which can be swept aside by the definition . . . of the military state. . . . It is at least as broad as the problem of civil-military relations. It exists in all states . . . . Thus, in our move from the ultraminimal to the military state, the problem of loyalty changes form—from external to internal, from private to public—but it does not disappear. Rather than a stark choice between treason and loyalty, we confront a relative comparison of risks.

Lastly, this formulation relates the link between demand and supply to one of the most fundamental dilemmas of liberal thought—the problem of how, in a world of rationality, self-interest, and scarcity, . . . legitimacy tames violence. . . . By definition, the MO wields a monopoly of the supply of military force. It is dominant. If the members of the MO are rational and self-interested, like everyone else, then why would they be loyal to the GO? . . . In answering these conundrums, what difference is produced by the distinction between a “private” or “public” MO? In other words, why would the MO of the military state be any more loyal than the MO of the ultraminimal state?

3. The Solution: A Culture of Loyalty

This is the central question of this Article, the main challenge to the dictum that “the state must monopolize force.” In response, . . . I expose the hidden sense in which our obvious answer to the problem of loyalty was essentially correct. I argue that in the most deep and direct sense, the “public” aspect of the MO—the idea that the MO is part of “the state”—is no more than a set of beliefs, symbols, and rituals that produce a kind of institutional wordplay themselves. When we say that our military institutions are “public,” we describe a culture of loyalty within those institutions.

There are two ways to make this point: We can reason from the culture of loyalty to the military state and from the military state to the culture of loyalty. Both ways highlight different aspects of the argument. . . .
a. Take one: From the culture of loyalty to the military state

Recall that in modern, liberal societies, we use three basic metaphors to describe organizations: “state,” “market,” and “community.” Each one depends upon a different principle to organize and motivate collective activities: “force,” “profit,” and “duty.” Let us apply each of these principles to the problem of loyalty and see which one fits best.

i. Force

It should be instantly clear that the political and legal principle of *force* cannot ensure the loyalty of the MO. Political and legal reforms do not leap off pages to punish military betrayals. They must be enforced. But in physical terms, the GO is impotent and the MO is dominant. If the GO entreats or hires helpers, the MO would defeat those helpers. The GO cannot force the MO to do anything.

ii. Profit

The economic principle of *profit* is only slightly more interesting: Perhaps . . . loyalty . . . could be ensured by the prospect of self-interested, reciprocal exchange . . . . Perhaps if the GO collects enough taxes, it could purchase the loyalty of the MO. Perhaps if the MO betrays the GO, the GO could withhold payment, or boycott the MO. The GO could ensure loyalty through economic reforms: contracts, inspections, and redistributions.

These economic hopes sound familiar, but they do not sound plausible . . . . They are premised on the GO’s ability to collect and protect taxpayer funds and enforce economic reforms. Let us take these three premises in reverse order: economic enforcement, tax protection, and tax collection. *Enforcement* still does not work. Like political and legal reforms, contracts, inspections, and redistributions do not enforce themselves. In the market, the GO is still impotent and the MO is still dominant. *Protection* is only mildly more plausible. If the MO wants tax dollars, it need not earn them by obeying orders. It can just take them—from politicians, judges, and administrators, or from the people themselves. Even *collection* fails. . . . How could the GO collect taxes? . . . If private policing firms enforced the state’s tax codes, then who would police them? More importantly, who would protect them? They would be robbed and extorted by the MO.

iii. Duty

This leaves the principle of *duty*, which turns out to be the ace up our sleeve. . . . As I explained earlier, the principle of duty is neither corporal nor capital, but cultural. Duty represents all of our voluntary, non-material modes of organization and motivation. Duty is our collective conscience—our senses of sanctity, morality, legitimacy, and justice. Our duties are created and sustained by our systems of beliefs, symbols, and rituals.

Why does duty work best in the military state? Force and profit assault the MO’s self-interest on the surface, in the bodies and wallets of generals, sergeants, and soldiers. But duty assaults self-interest at the core, in the “minds,” “hearts,” or “souls” of generals, sergeants, and soldiers. It does not work against the MO’s self-interest, like force. It does not work with the MO’s self-interest, like profit. It works through the MO’s self-interest.
It transforms and transcends the MO’s self-interest altogether. It fuses individual and institutional interests.

Of the three principles, duty is the only one which neither presumes nor requires any threats or acts of force. It requires only an act of union: In one way or another, the MO comes to consider itself bound to the GO and, thus, obliged to the GO. Together they form a single “community,” like a family, a neighborhood, a religion . . . or a state. Generals, sergeants, and soldiers come to adopt a system of beliefs, symbols, and rituals that ties them to the mast of the state . . . . They come to respect, obey, and enforce the decisions of the state . . . for the sake of the state itself. . . . [T]hey endorse the authority and legitimacy of the state. In short, they develop a military *culture of loyalty*.

The absence of force in this process is the key point. In the encounter between the GO and the MO, the GO stands naked before the MO. There is no third party to enforce regulations, contracts, inspections, or redistributions. If the MO betrays the GO, the GO is instantly lost. In this encounter, the state is created, and thus, it cannot be presumed. . . . [T]his is a unique situation. It only occurs in the institutional relationship between the GO, which monopolizes the demand for military force, and the MO, which monopolizes the supply of military force. . . .

Like the analogous concepts of “deliberation,” “authority,” and “legitimacy,” the culture of loyalty achieves a remarkable feat. It reduces one of the most fundamental tensions of liberal thought. . . . [T]he liberal state’s destiny turns on this military culture of loyalty.

iv. Duty, loyalty, and monopoly

But terms like “force,” “profit,” and “duty” are relentlessly abstract. Like “state,” “market,” and “community,” they are ideal types. In the real world, institutional relationships depend upon complex collisions of force, profit, and duty in everyday settings. Thus, . . . the problem of loyalty never disappears. In some institutional relationships, it is relatively large; in others, it is relatively small. But it always exists.

For this reason, I am not suggesting that everything is perfect and pure in the fantasy worlds of the “state of nature,” the “ultraminimal state,” and the “military state.” . . . In both the ultraminimal and military states, the relationship between the GO and the MO would depend upon force, profit, and duty. In both states, the MO would be bound to the GO by political, legal, and economic reforms, and cultural norms. After all, shareholders, managers, and employees are not unadulterated individuals made up purely of egos and ids. They are voters; they are jurors; they are citizens. They have some sense of public duty, some sense of loyalty to the state.

But there is a difference: They are not generals, sergeants, or soldiers. For them, loyalty is more like a hobby or a job, and less like a vocation or a profession. They sign contracts, but they do not swear oaths. They do not eat state foods, wear state clothes, and sleep in state beds. In everyday life, they do not constantly hold beliefs, share symbols, and practice rituals that bond them to the state. In other words, private military corporations are not “public.”
In some ways, then, my central claim is quite modest. In the ultra-minimal state, the MO is private; in the military state, the MO is public. In all other ways, the two states are the same. My claim is that if there is any distinction between these two relationships, it must be some relative distinction between profit and duty, between capital and cultural connections. A private MO would tend to place profit above duty; a public MO would tend to place duty above profit. A private MO would tend to think of itself first and the state second. It would be more likely to betray the state. A public MO would tend to think of itself second and the state first. It would be more likely to obey the state.

In different scenarios, the difference created by the military state could range from the subtle to the severe. But the difference would always remain. Since it would always be a relative difference, the military state would always enjoy a relative advantage. But all other things being equal, it would be an advantage of utmost importance.

b. Take two: From the military state to the culture of loyalty

. . . [L]et us flip the argument around to reconsider it from behind. Whereas before we reasoned from loyalty to publicness, now we will reason from publicness to loyalty. A moment ago, we asked ourselves: Which would be more loyal to the state, a public or private MO? Now we will ask ourselves: In the military state, in what sense is the MO “public”? Which “public” qualities of the MO matter most? . . . [O]n the deepest level, in the most direct sense, what makes the [U.S.] Armed Forces “public” for us?

Surely, at different levels, the [U.S.] Armed Forces are public in many ways. First, the[y] have political and legal links to the . . . government. . . . [T]hey are bound by the Constitution and . . . laws . . . . They must obey the orders of the Commander-in-Chief and the Military Code of Justice. If generals, sergeants, and soldiers disobey . . . , they can expect to be reprimanded, dismissed, or punished. Second, the[y] have economic links to the . . . government. They are owned, managed, and staffed by the . . . government rather than by any particular politicians, judges, administrators, or citizens[, and] draw upon taxpayer funds . . . . Third, the[y] have cultural links to the . . . government. They consider themselves “parts” of the . . . government, just as legislatures, courts, and agencies do. They consider themselves one step removed from the market, like a special breed of civil servants or public professionals. They hold beliefs, share symbols, and practice rituals that place them aboard the ship of state.

At different levels, then, the metaphors of “public politics,” “public law,” and “public economics” can be . . . applied to the relationship between the [U.S.] government and the . . . Armed Forces. This is especially true of the relationship between the . . . government and individual generals, sergeants, and especially soldiers. But when we consider the . . . Armed Forces—not as groups of individuals, but as organizations, or a network of organizations—it becomes rather implausible to speak in political, legal, and economic terms. Ask yourself this: Why do the . . . Armed Forces obey the Commander-in-Chief or the Military Code of Justice? Why do the[y] ask Congress for budget increases?

In any other public encounter, the answer is clear: “Public politics,” “public law,” and “public economics” are metaphors backed up by threats and acts of military force. Upon
every other public relationship, the state’s monopoly of the supply of military force casts a shadow. . . . Corporations, associations, and individuals obey the state’s . . . decisions because if they did not . . . , they would become targets of force. This . . . threat of physical coercion is, in all other cases, the distinctive characteristic of those political, legal, and economic relations that we call “public.” In politics, law, and economics, public relations are mandatory (and collective), or they are not “public” relations at all.

But in the public encounter between the GO and the MO, there is no threat of force in the background. . . . [P]oliticians, judges, and administrators do not bully or buy the obedience of the . . . Armed Forces. . . . Here the concepts of public politics, public law, and public economics lose traction. But the concept of “public culture” holds fast. . . . [T]he . . . Armed Forces . . . are public because they think they are public. In the end, it is as simple as that.

4. Monopolization as Socialization: Ownership, Management, and Employment

By analyzing the state’s monopoly of military force as a culture of loyalty, we propose a link between the monopolization and socialization of military institutions in liberal states. . . . Now we may study this link more closely. Rather than thinking in terms of the relationship between “the military” and “the state,” we may think in terms of three relationships between the two entities: . . . public “ownership,” “management,” and “employment.” We move from “why” questions to “how” questions: How is monopolization related to socialization? When armies are monopolized, how are generals, sergeants, and soldiers socialized to obey the state? How do public ownership, management, and employment structures instill a sense of duty, or a culture of loyalty, in military institutions and personnel?

Complete answers to these questions . . . call for . . . a full foray into anthropolog[y] and sociolog[y] . . . . But the liberal theory of force is not anthropological or sociological. It is a decidedly philosophical, speculative endeavor. On these “how” questions, our speculations are drawn from two familiar sources: first, basic principles of public ownership, management, and employment, laid out in numerous works of modern political, economic, and social theory; second, common knowledge of modern military institutions and cultures. . . .

. . . [T]o link monopolization to socialization, we can relate [ownership, management, and employment] to three cultural concepts: identification, professionalism, and bureaucracy. The links are not neat; there are overlaps on all sides. But for our present purposes, they suffice.

a. Public ownership

Public ownership has two distinctive qualities: It is collective and mandatory. In public armies, every citizen owns a “stake” in the armed forces, and every stake is the same “size.” Citizens . . . cannot buy or sell stakes. These two qualities produce two well-known effects: On the one hand, public ownership embeds inefficiency and rent-seeking within our armed forces; on the other hand, public ownership embeds public purposes
within our armed forces. By dispersing stakes and minimizing competition, public ownership insulates our armed forces from market pressures.

This difference between public and private armed forces would surely be a matter of grades. In many markets, individuals are more or less insulated from market pressures by employment contracts: Corporations hire them to deliver results and follow instructions and judge them in terms of both profits and duties. But in states, whole institutions are insulated from market pressures. Public ownership imposes institutional walls. Within these walls, public modes of management and employment flourish.

. . . [P]ublic ownership . . . also creates public places, which are infused with public values. Public ownership signals an identification between military institutions and the project of government. Because generals, sergeants, and soldiers are sustained by public property, surrounded by public symbols, and engaged in public rituals, they come to consider themselves public professionals and public bureaucrats.

b. Public management and employment

These days, the terms “professional” and “bureaucrat” are deeply contested, and thus, poorly defined. . . . [E]veryone wants to be a “professional” and no one wants to be a “bureaucrat.” . . . [Moreover,] mundane terms like “professional” and “bureaucrat” are rarely applied to generals, sergeants, and soldiers. In the liberal theory of force, however, these terms must be applied to military work. They each link the “insulation” of military work to three traditional military values: service, hierarchy, and loyalty.

Military officers are the managers of public armies, and in many ways, they are professionals. They believe that military management is a public service. They understand their work as a public calling, career, commitment, service, or duty. . . . They believe that their work is extremely important and connected to broad social objectives. . . . They swear an oath to uphold and defend the state. They typically work full-time, are often “on call,” and often live where they work. . . . They consider themselves a distinct group of workers: distinct from civilians; distinct from soldiers.

Soldiers are the employees of public armies, and in many ways, they are bureaucrats. Like officers, they are insulated from market pressures and the public at large. They are hired, promoted, demoted, and fired on the basis of highly formalized policies and rules rather than revenues and costs. They swear oaths, work full-time, and imagine their work as a public commitment, service, or duty (although not as a permanent “calling” or “career,” as officers do). During training and active duty, soldiers are separated from society, family, and officers. More than most other public workers, they pledge themselves to a term of service, which they may not abandon or shirk without severe consequences. As employees, soldiers are rigorously and meticulously disciplined and trained. To unique degrees, they are socialized to sacrifice self-interest in the name of military service, hierarchy, and loyalty. They are trained to exalt hierarchy and procedure, to prioritize obedience over ingenuity, and routine over results. They are inundated with symbols and rituals of military service, hierarchy, and loyalty: flags, emblems, uniforms, medals, slogans, and salutes.
5. Military Conscription: Professionals vs. Bureaucrats

When liberal states draft soldiers, the culture of military loyalty becomes especially strong and clear. In fact, the military draft is the most “public” aspect of public military employment. In other public institutions, ownership is always collective and mandatory, but in the military draft, public employment becomes collective and mandatory as well. In liberal states, other instances of collective, mandatory employment are exceptionally rare. Military conscription transforms military work into an actual act of public service . . . [and] demands a profound sacrifice of self-interest . . . .

. . . [T]he military draft highlights three important points. First, it demonstrates the different roles of officers and soldiers in the production and maintenance of a loyal armed forces. Second, it marks the limits of the “public” by showing us what we do not mean when we say that armies are “public” suppliers of force. Third, it points to another example of the tension between deliberation and domination: an example that occurs within military organizations themselves.

One simple observation helps to demonstrate these three points: In most liberal states, soldiers are drafted at times, but officers are not drafted at all. The two principles of the draft—collectivity and compulsion—are applied easily and often to soldiers. But when we try to apply them to officers, they make little sense.

Soldiering is the quintessential “bureaucratic” task, easily amenable to collective and mandatory execution. Soldiers . . . follow orders. In the soldier’s work, collectivity and compulsion are facts of everyday life. Officers, by contrast, tend to act alone, as leaders. Leadership may neither be democratized nor compelled. . . .

The distinction between officers and soldiers is flexible yet fundamental. . . . [I]n most armies, most of the time, the “real” officers—i.e., the members of the permanent, elite officer corps—are not members of the masses, and not forced to serve. In professional settings, principles of collectivity and compulsion may not be applied. As I just mentioned, this distinction has three implications.

First, it reveals the critical role that officers play in the relationship between states and armies. Officers bear the fundamental responsibility for the socialization of armed forces. . . . Soldiers cannot socialize themselves. . . . There must be a permanent officer corps so that there can be permanent armies with permanent loyalties. . . . Officers stand “between” soldiers, on the one hand, and politicians, judges, and administrators, on the other. . . . They establish and sustain . . . an insulated domain in which loyal soldiers evolve. When citizens are drafted, they are inundated with flags, emblems, uniforms, medals, slogans, and salutes, until they are transformed into soldiers. . . . [O]fficers . . . perform that critical cultural task.

The draft also suggests a second point: that our concept of “public” supply only goes so far. In most circumstances, “public” supply requires only that ownership be collective and mandatory. In “public” schools, hospitals, police departments, and prisons, employees volunteer. But these institutions are still “public” because they are publicly
owned. In armies, by contrast, we sometimes draft soldiers. In these moments, public employment becomes collective and mandatory too. But we never extend the principles of “collectivity” and “compulsion” to public management . . .

Finally, the draft reveals, in the most fundamental way, how deeply the tension between deliberation and domination actually runs. Earlier, we teased out the tension between the GO and the MO. Now we can see that the same tension exists within the MO as well . . . Generally speaking, soldiers use force more often than officers. In the most literal sense, it is more often soldiers than officers who “dominate” individuals and institutions, by carrying guns, driving tanks, flying planes, and above all, by aiming and firing weapons.

. . . Why do soldiers obey officers? Why not take over armies and states themselves? There are ultimately no answers to these questions . . . They show us that in the deepest sense, the liberal state’s monopoly of force always already presumes cultural links. Not just cultural links between the supply of force and the demand for force, which we have already seen, but links within the supply of force itself. In the absence of loyalty, there cannot be military institutions, let alone military states.

6. Market Socialization: Private Professions and Private Bureaucracies

With these observations in mind, we can return to the main line of the argument . . .: Service, hierarchy, and loyalty are the hallmarks of military life. . . . [I]t is most important to see the causal connection between insulation and socialization. Economic insulation is a necessary—although not remotely sufficient—condition for the creation and sustenance of extremely strong cultures of public service, hierarchy, and loyalty within institutions.

Two questions clarify this point. Why do other, non-military, public institutions lack such strong cultures of service, hierarchy, and loyalty? In other settings, the problem of loyalty is less vital and less severe. Treason costs less for states and earns less for traitors. Why do private corporations lack such strong cultures of service, hierarchy, and loyalty? In markets, these cultures cannot develop to the same degree. With less shelter from market pressures, and with less identification with public projects, private managers and employees remain more . . . tethered to self-interest.

This last sentence raises some prominent eyebrows. There are many examples of private professions and private bureaucracies. One wonders, then, whether the ultraminimal state could hire a highly professional, highly bureaucratic military corporation . . . But . . . [i]n the ultraminimal state, why would highly professional, highly bureaucratic military corporations exist? . . . Simply because politicians, judges, and administrators desired them? Because the state drafts contracts calling for “professionalism” and “bureaucracy” . . . ? Because the state solicits bids from only “professional” and “bureaucratic” suppliers? Even if such model corporations evolved, who would defend them against other, less scrupulous military corporations? Even if they could defend themselves, who would compel them to remain . . . professional [and] bureaucratic . . . ?

It is easy to imagine the ultraminimal state taking some significant steps. In most markets, there are ways to socialize corporations, by promoting the development of
professional and bureaucratic practices and norms. Most markets can be regulated. . . . [This] may work well in the professions of medicine and law, but in the military profession, they would be petty and out of place. The concentrated powers of military institutions create a concentrated danger of treason that calls for a concentrated form of socialization. In the relationship between the liberal state and its military institutions, the problem of loyalty peaks. In response, the liberal state must socialize generals, sergeants, and soldiers strictly. It must *monopolize* the supply of military force.

**D. Beyond the Military State?**

. . . In some ways, the story of the military state is much like the traditional story of the liberal state: there are individuals; there are fights; there are institutions. From conflict comes togetherness. In our story, we say that the principle of duty tames the principle of force, or more precisely, that the demand for force tames the supply of force. In the annals of liberal thought, these are familiar motifs . . . .

. . . [B]ut there is one glaring distinction between our story and others: The military state is much leaner than the traditional liberal state. . . . [It] relies completely upon private policing and punishment corporations. Why? From the traditional liberal perspective, this limitation makes little sense. After all, the military state is constructed to resolve the problem of loyalty. In the broadest sense, the problem of loyalty is boundless. It is like the problems of legitimacy, authority, and agency; it arises whenever one institution—indeed one individual—wields power over or depends upon another.

But ours is not a traditional liberal perspective. This is a liberal theory of force, not a liberal theory of morality or politics. The military state is not constructed to resolve every incidence of the problem of loyalty. It is constructed to cure only the most unusual, severe case: *the problem of order*.

What is the problem of order? In the ultraminimal state, the problem of loyalty reaches a climax. The prospects of disloyalty grow exceptionally high, and the consequences of disloyalty grow exceptionally grim. . . . [T]he ultraminimal state has too little ability to socialize private military corporations, and . . . private military corporations have too much ability to overthrow the ultraminimal state. As a result, disloyalty is linked to disorder. The ultraminimal state is unstable. . . . It cannot establish and maintain *order*.

. . . [T]he problem of order justifies the formation of the military state. . . . [T]he liberal state must monopolize the supply of force so that it can establish and maintain order. This is the first thesis of the liberal theory of force, which we will call the “principle of order.” In the last section, we identified and justified the principle itself. In this section, we identify and justify the limits of the principle. . . .: How far does the principle of order go? How much of the supply of force must the liberal state monopolize? The principle itself suggests an answer: The liberal state must monopolize enough of the supply of force to establish and maintain “order.”

But how much is that? Does the principle of order justify anything beyond the military state? Does it justify a monopoly of policing? A monopoly of punishment? If not, are
there other, more extensive principles that might? In particular, given that there is a principle of order, is there also a principle of law?

1. The Principle of Order . . .

. . . [T]o identify the limits of the principle of order, we need a . . . deeper understanding of the principle itself, and the problem that the principle cures. . . . [W]e need to hone in on the link between disloyalty and disorder. . . . [D]isorder is a much more finite problem than disloyalty. Almost any institution can betray the state, but very few institutions can topple it. . . . [W]e cannot say . . . that the private prison industry [can] subjugate the state. Prisons can commit crimes and violate the state’s laws, but they cannot wage wars and upset the state’s order. Armies, by contrast, can do all of the above.

Why? Why, in relative terms, are armies so dangerous and prisons so docile? Why is military force such a singular form of force? Why is military disloyalty so strongly linked to disorder? The answer is intuitively clear: Military force is the most intense form of force. . . .

. . . [T]he problems of order and loyalty . . . are both linked to the intensity of force. . . . Where the intensity of force is greatest, the problem of loyalty is greatest, and the problem of order exists. Where the intensity of force is least, the problem of loyalty is least, and the problem of order does not exist. Perhaps pictures are clearer than words:

**FIGURE 5**

. . . Like the demand/supply matrix, the loyalty/intensity graph is a metaphor. Please do not take it too literally. In particular, please do not infer anything from the shape or the slope of the curve, or the evenly-spaced arrangement of “punishment,” “policing,” and “military force” along the X-axis. . . .

But one feature of the graph is fixed: The problem of loyalty cannot reach “zero” in the case of punishment. In punishment institutions, the problem of loyalty diminishes, but it
does not disappear. The curve depicts relative comparisons among the three forms of force, rather than absolute descriptions of each one. . . . [T]he problem of loyalty is greatest in the case of military force, and least in the case of punishment force. . . .

2. . . . and the Intensity of Force

. . . What is “intensity”? Why is policing more “intense” than punishment? . . . “Intensity” breaks down into . . . “severity,” “breadth,” and “speed.” Severity is a measurement of the physical harm that any exercise of force does . . . . Breadth is a measurement of how widely that harm is distributed . . . . Speed is a measurement of how quickly that harm occurs. Intensity is the quasi-mathematical “product” of all three elements.

. . . Each element bespeaks a liberal danger that inheres in any exercise of force: the danger [of] . . . illegitimate[ ] harm[] . . . [W]e are not talking about mistakes or accidents here. . . . We are talking about the danger of disloyalty, which involves the danger of intentional illegitimate harms.

. . . We can explain “legitimacy” in terms of allocations of force and “harm” in terms of exercises of force. Targets and levels of force are “legitimate” when they are “legitimized” by the liberal state’s allocative bodies, through liberal political and legal procedures. “Harms” are the physical effects caused by exercises of . . . force. Because the term “exercise” includes threats, the term “harm” includes threat-based deprivations, such as physical confinement and the confiscation of property. . . .

a. Severity

Surprisingly, “severity” turns out to be the most puzzling element. On one level, it seems clear that punishment is the minimum (i.e., less severe than policing and military force); on another level, it seems clear that military force is the maximum (i.e., more severe than policing and punishment). But there is no single level of analysis in which we can consistently distinguish the severity of all three forms of force.

The first level is the abstract/possible mode of analysis. It compares the severity of punishment, policing, and military force in the abstract. For example: In the United States, our public armies, police, and prisons all kill, hurt, and detain people, and destroy, damage, and confiscate property. From this perspective, severity is constant in the United States; American practices of punishment, policing, and military force are equally severe. But in this respect, the United States is not a representative sample of liberal states around the world. In most other liberal states, prisoners are not executed; capital punishment is no longer permitted. So in most liberal states, punishment is less severe than policing and military force. In other words, punishment is the “minimum,” and policing and military force share the “maximum” together.

The second level is the actual mode of analysis, which is . . . more . . . complex. To see the actual level, we must compare the “average” severity of punishment, policing, and military force. . . . [We] compare[] the most common means of physical coercion in liberal punishment, policing, and military institutions.
. . . In these actual terms, military force is clearly the “maximum”; it is obviously more severe than policing and punishment. . . .

The puzzle pops up when we try to compare the “actual” severity of policing and punishment. What should we compare? In coercive terms, what are the standard operating techniques of policing and punishment? Is policing best represented by search, arrest, or detainment? Is punishment best represented by parole or imprisonment? . . .

. . . I do not think that these questions have any sensible answers. . . . So for our purposes, let us conclude that . . . in terms of severity, military force is the “maximum,” and policing and punishment share the “minimum.” . . . As I argue below, the distinction between policing and punishment is more aptly conceived in terms of breadth and speed.

b. Breadth

. . . [B]readth . . . is relatively easy to apply and adds much definition to the three forms of force: Military force is the most broad; punishment is the most narrow; policing falls between those two extremes. . . .

We can conceptualize the element of “breadth” in two ways—by considering particular exercises of force and by considering the “aggregate” of all exercises of force. On the particular level: Liberal states almost always exercise punishment on an individual basis. Liberal states typically exercise policing much more broadly to suit the exigency of policing work. When circumstances warrant, police arrest multiple individuals, search multiple locations, and seize multiple things. Lastly, liberal states often exercise military force on much broader scales . . . .

In some sense, however, it is . . . arbitrary to speak of particular “exercises” of force. To meet this objection, we can speak in aggregate terms: The total harm of punishment is roughly the number of people who are currently in prison, and, perhaps, on parole. In the United States, this harm is unusually broad, with roughly two million people in prisons or jails and another five million people on probation or parole. . . .

Now there are many different ways to calculate the population of policing, but I do not want to get too caught up in them here. Whether we say that all citizens are policed, or that only suspects are policed, the policing population is . . . greater than the punishment population. Every person in prison was probably a target of policing at some point, and many people who are not in prison are obviously targets of policing as well.

Compared to punishment and policing, the scope of military force is much broader still. Once we remind ourselves to include threats of military force, we instantly see that the targets of military force cover the globe. . . .

To sum up: In terms of breadth, punishment is the “minimum,” policing is the “median,” and military force is the “maximum.”
c. Speed

Lastly, there is the element of speed (which might also be called “haste”). Speed distinguishes punishment very clearly from policing and military force, but it does not significantly distinguish policing from military force. In relative terms, exercises of punishment are slow, while exercises of policing and military force are quick.

In some ways, this is the simplest distinction yet. It is produced by liberal political and legal procedures. More specifically, it is produced by the liberal criminal justice system, in which prosecutors, judges, and juries select the targets and levels of punishment on a case-by-case basis—before punishments are actually inflicted.

. . . [T]his is an abstract . . . distinction, and in actual terms, there are [exceptions] . . . . But these are quibbles. In relative terms, the temporal distinction is clear: In liberal states, policing and military force are exercised quicker or more hastily than punishment; punishment is exercised slower or more deliberately than policing and military force.

On average, the procedural constraints placed upon exercises of policing and military force are more likely to be ex post than ex ante, weaker than stronger, and more general than specific. Police and military personnel often choose particular targets and levels of force by themselves . . . because they must make graver decisions quicker, more flexibly, and more frequently than prison personnel. Policing and military force are typically exercised in response to more urgent and changing concerns . . . . By contrast, the goals of punishment . . . are typically regarded as less pressing concerns. If our punishment wheels grind slow and fine, then our policing and military wheels spin fast and rough.

So in terms of speed or haste, punishment is clearly the “minimum.” Is there a temporal distinction between policing and military force? There might be, but once again, I find myself unable to articulate one in any coherent, consistent terms. . . . So let us say that . . . punishment is the “minimum” and policing and military force share the “maximum.”

d. Summary

[Let us] complete our quasi-multiplication tables. This should convey what I mean by the “intensity” of force, and why I think that punishment, policing, and military force are, respectively, the least, median, and most intense forms of force.

<table>
<thead>
<tr>
<th></th>
<th>Severity</th>
<th>Breadth</th>
<th>Speed</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Military</td>
<td>Maximum</td>
<td>Maximum</td>
<td>Maximum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Policing</td>
<td>Minimum</td>
<td>Median</td>
<td>Maximum</td>
<td>Median</td>
</tr>
<tr>
<td>Punishment</td>
<td>Minimum</td>
<td>Minimum</td>
<td>Minimum</td>
<td>Minimum</td>
</tr>
</tbody>
</table>

FIGURE 6

660
Draft — do not distribute!
Before we move on, one last observation on the topic of intensity might be useful: In liberal states, every exercise of punishment, policing, and military force must be . . . proportional to [threatened or actual harm]. The concept of intensity could also be worked out and applied in terms of these harms. Rather than speaking of the intensity of legitimate force, we could speak of the intensity of illegitimate violence: unpunished crimes, unpoliced crimes, and military threats. Broadly speaking, our conclusions should be the same: Liberal states should act more severely, broadly, and quickly in response to more severe, broad, and urgent threats and harms.

3. Intensity, Supply, and Demand

. . . In our search for the limits of the principle of order, we are juggling two concepts: the intensity of force and the problem of loyalty. . . . [W]e are developing a . . . relationship between the intensity of force and the danger of disloyalty: Where the intensity of force is greatest, the danger of disloyalty is greatest[, and vice versa]. . . . We can say that the intensity of force “causes” the danger of disloyalty.

The statement is . . . not very satisfying. It does not tell us much about why and how intensity could “cause” disloyalty. As a result, it leaves us unable to pinpoint where on the X-axis the problem of loyalty is substantially diminished, such that the problem of order does not exist. In order to identify the limits of the principle of order, we must fully understand the sense in which intensity “causes” disloyalty and disorder.

. . . [W]e must translate the terms of “intensity” and “disloyalty” into more familiar . . . idioms. Returning to the familiar discourse of supply and demand is a helpful start. We can readily define “intensity” in terms of the supply of force: Intensity tells us how severely, broadly, and quickly . . . force is exercised. Then we can define “loyalty” in terms of the relationship between demand and supply: When loyalty exists, the payment for force induces the faithful provision of force; the allocation of force induces the faithful exercise of force. When disloyalty exists, then supply does not satisfy demand. Simply put, armies, police, and prisons do not do what states ask them to do.

. . . [W]hen we say that intensity “causes” disloyalty, we mean that where . . . intensity . . . is greatest, the link between demand and supply is weakest. . . . [W]here . . . intensity . . . is greatest, those institutions that finance and allocate force have less control over those institutions that provide and exercise force[;] the power of supply is most concentrated; it threatens to subsume the power of demand.

. . . When we first introduced the ultraminimal state, we expressed . . . the same concern: We said that in the state of nature, the only organization that could finance and allocate force . . . must also be . . . an organization that could provide and exercise force. Next, when we compared the ultraminimal state to the military state, we expressed a secondary version of the same concern: We noticed that the MO was uniquely positioned to betray the GO. This third version is the last. Now we can link the intensity of force to the inherent tension between supply and demand. We can describe this link systematically and explicitly in causal terms.
When we form an organization[, like an army,] that wields th[e] . . . ability to inflict harm severely, broadly, and quickly, we form an organization that enjoys a singular opportunity—and thus, experiences a singular temptation—to subjugate the state.

Fortunately . . . public and private armies are not equally tempted to capitalize upon this unique stroke of luck. But tragically, all other things being equal, public and private armies are equally equipped to betray the state. Thus, when we monopolize the supply of military force, we minimize the danger of disloyalty, but we do not altogether avoid it. The concentration of power inherently and inevitably produces the opportunity of abuse, and to some degree, it inevitably produces the temptation of abuse as well.

Thus, on our earlier graph, the intensity/disloyalty correlation was not accidental. In comparative terms, the intensity of force “causes” the danger of disloyalty in two different senses: (1) Intensity is the reason that the stakes of military disloyalty are higher (the liberal state could fall, and the greatest amount of life, liberty, and property could be lost); and therefore (2) it is also the reason that the risk of military disloyalty is higher, all other things being equal. Of course, intensity is not the only cause of disloyalty, but it is a contributing factor, and in the context of military institutions, it is an exceptionally powerful one. When treason is easier to commit and likelier to succeed, traitors are easier to find and more likely to act. We should therefore expect that the danger of disloyalty will be greater when, where, and because force is more intense.

4. The Governability of Force

To clarify our thesis, it is helpful to slap a more familiar label on this variable dynamic between supply and demand. Let us call it the “governability” of force. . . . [W]here the intensity of force is greatest, the governability of force is least; and where the intensity of force is least, the governability of force is greatest. In other words, we may picture an inverse relationship between the governability and intensity of force:

**Figure 7**

Draft — do not distribute!
More questions arise: What is the “governability” of force? Is “governability” the same as “loyalty”? The two concepts are related, but governability is not simply loyalty under a new name. It is the ability of states to promote loyalty through political, legal, economic, and cultural reforms. It is the ability of demand institutions to control supply institutions.

As we have already suggested, and as the graph depicts, liberal states are least capable of governing exercises of military force. Because armies wield the greatest ability to inflict legitimate harm, they necessarily enjoy the greatest opportunity to inflict illegitimate harm. Thus, they necessarily experience the greatest temptation to inflict illegitimate harm . . . . The supply of military force always threatens to deviate from the demand for military force. If it is true, it is truly tragic. It vindicates the view that power tends to corrupt, and more specifically, that the most intense forms of power tend to corrupt most intensely. Let us briefly consider the idea from political/legal and economic perspectives.

\[ a. \text{Allocation and exercise} \]

In the political/legal idiom, the governability graph says: Where the intensity of force is highest, institutions that allocate force have less ability to control institutions that exercise force. Politicians, judges, and administrators cannot control exercises of military force through political, legal, and economic reforms alone because they cannot enforce redistributions, contracts, inspections, and regulations against military institutions. . . . To support such reforms, politicians, judges, and administrators must also resort to “cultural” controls. . . . They must instill a culture of loyalty in military institutions because military institutions exercise the most intense forms of force. . . .

Yet the graph suggests another point, which turns out to be a strange silver lining for minimalists. For the moment, let us presume that a military state emerges, so that the state does monopolize the most intense forms of force. Now the balance of power shifts . . . . The danger of disorder recedes. Now officials can rely on public armies to enforce [the law] against other institutions. The need for further monopolization apparently fades.

\[ b. \text{Payment and provision} \]

In economic translation, the governability graph says: Where the intensity of force is highest, institutions that finance force have less ability to control institutions that provide force. This is not mere repetition. As I said earlier, public allocations of force beget public payments for force. . . . [P]ublic spending is a rough reflection of our collective choices regarding when, how, how much, and against whom force should be used. When Americans say that they want to topple the Taliban, close down our borders, and throw terrorists in jail, these demands are eventually reflected in our public fisc.

In economic terms, the governability graph proposes two different ideas, which are both related to this basic link between public payment and public choice. First, it suggests that the purchasing power of public funds—the proverbial bang for our buck—is inversely related to the intensity of force. Where the intensity of force is greater, officials must spend more tax dollars to successfully control exercises of force; in other words, they receive less control for each tax dollar that they spend. As a result, budgets swell.
Second, and similarly, the graph proposes that where the intensity of force is greater, tax dollars are more effectively spent upon “monopolization” rather than other reforms, such as redistributions, contracts, inspections, and regulations.

But there is a bright side in economic terms too: Once the state monopolizes military force, the economic tide turns. Vis-à-vis other institutions, the purchasing power of public funds becomes greater. It is now more effective to invest in political, legal, and economic reforms rather than investing in cultural ones.

5. Governability vs. Efficiency

Note the re-emergence of economic notions in terms like “purchasing power” and “effective spending.” These are not only metaphors now. . . . It is an instructive moment.

. . . [I]n our public policy debates, the efficiency argument was our permanent pest. . . . [P]olitical principles [were] always subject to complex and indeterminate economic tradeoffs. Possible gains in “accountability,” “rights,” or “distributions” must be traded against one another and weighed against possible losses in “security,” “protection,” “safety,” and, most of all, “efficiency.” It was difficult to say, with any degree of certainty, that the efficiency argument could be safely ignored.

In the ultraminimal state things became clearer. Resources were still scarce so the efficiency principle still generally applied. But where the intensity of force was greatest—in private military corporations—the principle proved nothing. In the ultraminimal state, the efficiency argument fell apart. Politicians, judges, and administrators had no reliable way to sustain a link between the demand for military force and the supply of military force. . . . [T]hey had no way to govern military force at all.

In this atmosphere, it was naïve to argue that private armies would be more “efficient” . . . . The risk was too great . . . that private armies would be more efficient at rebellion . . . than at protecting the state. In the pursuit of profit, they would reject the state’s contracts and . . . enter the business of government for themselves. . . . [T]ax dollars were more likely to pay extortion fees . . . than to buy more efficient battle plans and victories. In the ultraminimal state, the efficiency argument was clearly unsound.

But now, as we . . . step away from the ultraminimal state, the efficiency argument regains some ground. Once the state monopolizes armies, the link between demand and supply becomes more stable, and the problem of loyalty becomes more subdued. . . . [T]he consequences spill over into other institutions and industries. The military state may back redistributions, contracts, inspections, and regulations with threats and acts of force. Means of governance proliferate. It suddenly becomes plausible for politicians, judges, and administrators to think, talk, and act more broadly in efficiency terms.

6. The Principle of Law?

Where the intensity of force is least, the governability of force is greatest. This relationship generates a surprising result: In the liberal theory of force, the principle of law is weak and contingent. In contrast to the problem of order, the “problem of law”
presents less severe, broad, and urgent harms. As a result, law enforcement institutions are more amenable to political, legal, and economic reforms. They require weaker cultural controls. In lieu of monopolies, the liberal state may use redistributions, contracts, inspections, and regulations to promote loyal law enforcement. . . . The liberal state must monopolize the maintenance of order, but it need not monopolize the enforcement of law. But this entails some deeply controversial conclusions, which must be justified and explained.

\[a. \text{The principle of punishment}\]

We can start with the contrast between armies and prisons, which provides the clearest contrast between order and law. Here my conclusion is straightforward: Once the state monopolizes armies, it no longer needs to monopolize prisons. In the military state, [there are] other ways to promote loyalty in prisons. In lieu of monopolization, [one] can impose political, legal, and economic reforms upon private prisons. [One] can rely exclusively upon redistributions, contracts, inspections, and regulations.

In the military state, would private prisons be disloyal? Would there be corruption, extortion, and abuse? Would some inmates be punished too much and others not punished enough? Of course. All institutions falter and some institutions fail. But in the military state, when private prisons are disloyal, all hope is not lost. Politicians, judges, and administrators can call upon public armies to impose loyalty upon private prisons. When they identify errant firms and employees, they can withhold payments and terminate contracts. They can throw one private prison’s guards in another private prison’s cells. Backed up by threats and (if necessary) acts of . . . force, the state can enforce redistributions, contracts, inspections, and regulations against private prisons.

Even if things happen to go completely awry, all is not yet lost: The military state can use public armies to “take over” private prisons temporarily until more loyal suppliers can be recruited and hired. . . .

. . . [Thus,] the risk of prison disloyalty is minimal in the military state. The military state would design and regulate the market for punishment. It would reward loyalty and punish betrayal through political, legal, and economic reforms. Private prisons would have strong incentives to fulfill orders. They would have weak incentives, and fairly few chances, to do anything else. On the whole, schemes of corruption, extortion, and abuse would yield few profits and involve substantial risks and costs.

If this picture is too rosy for you, it can be easily darkened. If you prefer, you may suppose that the market for punishment malfunctions. Suppose that private prisons do not deliver the services that liberal values demand. Corruption, extortion, and abuse become rampant in private prisons. Who would be to blame? The state or the market? The customers or the providers?

If we mean to produce a fundamental criticism of private prisons, we must answer these questions definitively and persuasively, in the fairest possible terms. But in the liberal framework, we cannot. The moment that we lay blame upon private prisons, we entertain
some possibility of market failure. In other words, we put ourselves back . . . into the heart of our public policy debates. We say, for example, that private prisons would create “agency” costs, or peddle “influence” with politicians and administrators.

More conceptually, when we make these critiques, we enter a world in which tradeoffs between loyalty and efficiency are inevitable and theoretical certainties are scarce. In [this] world, it is no longer enough to prove that private prisons would be disloyal. We must also prove: (1) that private prisons would be more disloyal than public prisons; (2) that the disloyalty of private prisons would be “worse,” in liberal terms, than the inefficiency of public prisons; and (3) that the disloyalty of private prisons could not be diminished by political, legal, and economic reforms. It makes little sense to entertain private doubts, unless we play fair by entertaining public doubts too.

Of course, these three propositions are neither intuitive nor obvious; they are deeply empirical and contextual. At the very least, they permit the military state to experiment with private prisons. In this way, these propositions are easily distinguishable from the argument against private armies. Private armies have an obvious and inherent ability to destroy political, legal, and economic structures; private prisons do not. In most respects, private prisons must take the market for punishment as they find it. In this regard, they are less like private armies and more like . . . hospitals, schools, and the like.

b. The principles of policing

What about policing disloyalty? In the military state, could the private police use more local, subtle forms of force—corruption, sabotage, or guerillaism—to destabilize the state itself? The answer turns largely on what we mean by the distinction between “policing” and “military” institutions or, from another perspective, “law” and “order.”

The term “policing” encompasses many things indeed. In most liberal states, the distinction between policing and military institutions is typically described as a distinction between “internal” and “external” exercises of force. We often say that policing is exercised against citizens who live inside the liberal state’s boundaries, whereas military force is exercised against foreigners who live in enemy states. In peaceful times such geographic distinctions are more or less sound. But in moments of disorder—hostage crises, mob riots, terrorist attacks, and civil wars—they tend to erode. In these circumstances, police behave much more like armies and armies intervene domestically themselves. Thus, the distinction between “internal” and “external” is variable and contingent. In some liberal states, like the United States, they are fairly firm; in others, like Italy, they are fairly flexible, when they exist at all. For our fundamental analysis, it is much better to stick to the intensity of force—more specifically, to the severity and breadth of policing and military force. As we have already seen, exercises of policing are typically less severe and broad than exercises of military force.

. . . [I]n conceptual terms, we can divide policing into two parts. On the “law” side, there might be patrols, investigations, and arrests; on the “order” side, there might be home invasions, hostage crises, and riot control.
The most severe and broad forms of policing must be public; the least severe and broad forms of policing may be private. Why? Think back to the military state. Public armies can do some things well, but they cannot do all things well in all places. Military force . . . is the classic blunt instrument. By definition, it is the most severe and broad form of force. As the United States learned in Vietnam, and the Soviet Union learned in Afghanistan, severity and breadth have both pros and cons. In the military state, if strong private police forces formed local guerilla bands, they could wreak significant havoc . . . . By contrast, weak private police forces—patrollers, investigators, and arrestors—wield more limited powers . . . . They can betray the state and engage in terrible abuses of individual rights. But they cannot subdue the state itself . . . . Like private punishment, private law enforcement is amenable to political, legal, and economic reforms.

In fact, this division of policing is nothing new. When we discussed the distinction between armies and prisons, it was already implicit in our solutions. We said that the military state could use public armies to impose loyalty upon private prisons by enforcing redistributions, contracts, inspections, and regulations against private prisons. This was true. But when public armies do such things, they would be capturing or confiscating more often than they would be killing or destroying. In other words, they would be acting as public police forces, not public armies. They would be “policing” private prisons . . . .

c. Strong and weak theories of force

There are actually two theories implicit in this argument, one strong and one weak. The strong theory of force, which I do not endorse, states that there is no liberal principle of law. Under the strong theory, the liberal state may only monopolize order and may not monopolize law at all. Public law enforcement is forbidden. It is not necessary; so it is not permitted.

Presumably, this theory is motivated by a radical suspicion of public powers, which falls just short of anarchism. In its strongest forms, it is not a suspicion that I share. As we saw earlier, private powers are very problematic, even in the modern administrative state. There are many grounds upon which we might justify a principle of law: accountability, democracy, and rights. These are relatively contingent, empirical, and contextual grounds, but they are worthy grounds nonetheless. After all, we live in a contingent, empirical, and contextual world.

The weak theory of force, which I do endorse, makes room for those practical realities. It is just like the strong theory of force, without any special suspicions of public powers. The weak theory states that there is a liberal principle of law, but it is not a fundamental principle of liberal government. Under the weak theory, the liberal state must monopolize order, but it may also monopolize law. The distinction remains, but it is weaker. The door remains open for more expansive states. Liberal states may experiment with private prisons and private police to promote values like accountability, democracy, and rights.

The weak theory of force has one distinctive advantage: It enables us to conceptualize “monopolization” in degrees rather than absolutes. As we saw earlier, the distinction
between “public” and “private” is less like an “on/off” switch and more like a “dimmer” switch. It is a continuum between two kinds of ownership, management, and employment styles. On the “order” side, we are more likely to find personnel who have been drafted and deeply socialized by the state. On the “law” side, we are more likely to find personnel who have been paid and weakly socialized by the state. The weak theory simply states that the former should be soldiers and, perhaps, SWAT teams, while the latter may be prison guards and, perhaps, traffic cops.

In a nutshell, that is the liberal theory of force: Liberal states must monopolize order, and they may monopolize law.

E. The Private Demand for Force

Actually, that last formulation is not precisely correct. It constitutes only the “supply-side” of the liberal theory of force. Until now, a dubious presumption has lurked behind all of our previous graphs, principles, and hypothetical “states”—the idea that liberal states must completely monopolize the demand for punishment, policing, and military force, in equal ways, to equal degrees. In both theoretical and practical terms, it is simply not true. There are broad exceptions that swallow the putative rule. Think of our individual rights to self-defense and self-help. More broadly, think of our right to purchase private policing, with whatever wealth we might possess.

Recall that in most liberal states, you can actually “shop” for policing. If you happen to think that certain people should be imprisoned, you cannot hire a private prison to imprison them. If you happen to think that certain states should be invaded, you cannot hire private armies to invade them. . . . But if you happen to think that certain people should be policed, you can hire private police to police them. . . . Why policing? What special qualities of policing entitle individuals, associations, and corporations to provide and purchase it? In liberal states, why may there be a private demand for policing but not a private demand for punishment or military force? We can start with a few obvious points: In the state of nature, individuals would surrender some authority to the liberal state, but they would keep most authority for themselves. In liberal states, there are many decisions that politicians, judges, and administrators could not and should not make. These decisions are made by individuals, associations, and corporations instead.

For our purposes, there is only one relevant group of decisions—the who, when, where, how, and how much of punishment, policing, and military force. Individuals would reserve the right to make some allocations of force but not others. The relevant questions are: “Which ones?” and “Why those?”

. . . [O]ur working hypothesis is that individuals would reserve only policing allocations, and only a subset of those. . . . [L]et us say that they would reserve only the “law” aspects of policing allocations—those aspects that deal with prevention, investigation, and apprehension. They would not reserve the “order” aspects of policing, which involve home invasions, hostage situations, terrorist attacks, and riot control.
Most precisely, let us say that individuals would reserve limited rights to self-defense and self-help, as well as the limited right to hire private police, in order to aid the self-defense and self-help and others. They would keep no rights to hire private prisons or private armies, but they would keep a limited right to hire private police. These are completely conventional claims. But they beg the big question: Why those rights and not others?

There are many ways to justify these distinctions. We have already touched upon more than a few: property, privacy, accountability, equality, democracy, and rights. Yet the elements of the intensity of force are surprisingly helpful here too.

By hypothesis, let us say that individuals would reserve the right to inflict legitimate harms that are maximally quick yet minimally severe. . . . [L]ook[ing] back . . . , we . . . see that these are the same harms to which policing responds.

Is the hypothesis correct? I think so. It integrates many of the classic justifications for self-defense, self-help, and private policing.

1. Speed: Punishment vs. Policing

In the state of nature, individuals would surrender the right to allocate punishment, but they would keep a limited right to allocate policing. Why? Speed clearly has something to do with it. Every liberal definition of self-defense rights refers, in one way or another, to “imminent” harms. When harms are imminent, we may defend ourselves against them. When they are not we must forebear or call the public police. It is not just a question of supply. We may decide, within boundaries, how to use force, how much force to use, and against whom force is used. Our most drastic decisions are often reviewed by public authorities. But in the first instance, we allocate policing force ourselves.

By contrast, there is little rush to make punishment decisions. . . . There is plenty of time to ask politicians, judges, and administrators to intervene. . . .

2. Severity: Policing vs. Military Force

In the state of nature, individuals would surrender the right to allocate military force, but they would keep a limited right to allocate policing. Why? Severity clearly has something to do with it. Every liberal definition of self-defense rights refers . . . to “proportional” responses. When threats are minimally severe, we must defend ourselves in limited ways.

Of course, our responses need not be exactly proportional. We can use enough force to parry illegitimate threats. But the point is that in strong liberal states—states that monopolize armies—illegitimate threats to individuals are rarely severe. Individuals may elect to use bats, knives, or guns, but not bombs, tanks, or planes. In liberal states, the former are not necessary; so they are not permitted. . . .

3. The Space of the Self: Public and Private Domains

As we have already seen, breadth was the most useful way to distinguish between the supply of punishment, policing, and military force. After making several tries, I have
decided that breadth is not a useful way to describe the distinctions between the demand for punishment, policing, and military force. But there are good analogies—the concepts of public and private “spaces,” “places,” or “domains.”

We are all familiar with the liberal notion that politicians, judges, and administrators should not rule certain private domains. . . . It is a fundamentally moral distinction: These domains circumscribe individual rights.

These moral distinctions have strong implications for the private demand for force. Punishment and military decisions . . . can be . . . sequestered in public domains. Self-defense and self-help decisions—i.e., “self-policing” decisions—cannot be contained and removed in these ways. No matter how hard liberal states try, crime continues to force itself into private domains. Criminals enter businesses, communities, and homes, and individuals must fend for themselves. They may call the police . . . , but they need not ask for permission. If they are willing and able, they may act.

It is important to notice the way that technology and morality come together here. On the one hand, it seems that liberal states simply lack the capacity to be all things to all individuals, in all places, at all times. The public police cannot follow us wherever we go, deterring and fending off attacks. Thus, we keep our rights to self-defense and self-help, and, if these are not enough for our tastes, we keep our right to hire private police.

But it is not simply a matter of technological capabilities. Even if liberal states could police everyone, they would not. In moral terms, there are simply some places that liberal states may not be sovereign. They may not install a public police officer in my living room . . . . In liberal states, there are individuals, and thus there are “selves.” And wherever there are selves, there are rights to self-defense, self-help, and private police.

4. Conclusions on Private Demand

Of course, there are public spaces in liberal states too. This highlights the inevitable tension between public and private policing in liberal states. On the one hand, individuals, associations, and corporations provide and purchase policing. On the other hand, the state provides and finances policing. But as individuals associate, as private property expands, as the demand for policing expands, private policing corporations proliferate. They threaten to supplant public police departments, to enforce private “laws” upon hapless citizens. . . . [W]hen private policing markets grow and spread, our reliance upon private police can produce strange distributions and externalities. In liberal states, they must be allowed to exist, but like all private projects, they must be constrained and contained. . . .

This, then, is a vague sketch of the liberal theory of private demand. Of course, speed, severity, and space are not the only grounds upon which we could construct a liberal theory of the demand for punishment, policing, and military force. . . . As always, our public policy arguments also play a significant role. But here, as before, they work within limits. They do a decent job justifying a public demand for force, in general terms, but they do not effectively distinguish between punishment, policing, and military force.
Everyone wants prisons, police, and armies that are “accountable,” “democratic,” “egalitarian,” and “libertarian,” if we paint those values with broad enough strokes. But why, then, is there a private demand for policing but not punishment and military force? We clearly need a different set of arguments here. Speed, severity, and space are my preliminary thoughts.

**CONCLUSION**

So there you have it—the liberal theory of force. Out of the state of nature emerges a complex liberal state. It is neither an ultraminimal state nor a monopoly state. It has a public demand for punishment, policing, and military force, and a private demand for policing, in the service of self. It has a public supply of military force and a private supply of punishment. It has a supply of policing that is both public and private.

We can state the liberal theory of force in two theses. On the supply-side, there is the “intensity” thesis: In the name of loyalty, liberal states must monopolize order, and they may monopolize law. On the demand-side, there is the “self-policing” thesis: Liberal states must monopolize the demand for prisons and armies—and, to some degree, police forces—but they must also recognize the private demand for policing within private domains.

**A. The Problem of Order vs. The Problem of Law**

In philosophical circles, the most distinctive aspect of the liberal theory of force is the focus on “order” before, over, and above “law.” In most state-of-nature stories, the threat of crime takes center stage, and little attention is left for the threats of rebellion, civil war, conquest, and colonization. Such stories make short shrift of the problem of order and move directly to the problem of law. . . . The typical tale starts with the state of nature and jumps directly to the so-called “minimal” or “night-watchman” state, in which punishment, policing, and military force are all monopolized . . . .

Why have I bothered to be different? . . . I find the dearth of military analyses in liberal texts preposterous. In historical terms, the move from “order” to “law” represents a colossal leap in institutional development. Most of today’s modern, liberal states began to monopolize military force long before they began to monopolize policing and punishment. It is easy to guess why: It is not crime, but war that most often makes and breaks states. . . . To historians, this much is obvious; to laypersons, this much is intuitive. But in the state-of-nature tradition, this much is routinely and blithely ignored.

So philosophers are not historians. Fair enough. But even in philosophical terms, our military myopia is rather mysterious. . . . If individuals fight with one another, then why not institutions and states? If we need legislatures, courts, prisons, and police forces, then why not commanders, armies, navies, and air forces as well? Aside from the idiosyncrasies of liberal thought and liberal thinkers, there is no apparent method to this philosophical madness. In the state-of-nature tradition, the problem of law has been practically fetishized, but the problem of order has been badly neglected.
Of course, it is easy to offer excuses. Order is clearly not the ultimate goal of the liberal state. We might say that we have often skipped over the problem of order in our haste, in order to consider the more “liberal” projects of government like the creation, interpretation, and enforcement of laws, and, arguably, elections, schools, housing, health care, and welfare. But whether we like it or not, order is the first step on the liberal path. . . . A state must survive before it can thrive; it must be a state before it can be liberal. Individual rights cannot flourish in conditions of rebellion, civil war, conquest, and colonization. When the liberal state produces order, it lays the foundation for law. Law, in turn, lays the foundation for all manner of liberal projects.

In our story, then, first things come first. We do not skip off toward the liberal horizon, looking for problems to solve. We start with the most fundamental problem of government—the problem of order. We move to the most fundamental solution—the military state. Step by step, we build a liberal theory of force. What we find, in the end, is a liberal surprise—a state that is neither “minimalist” nor “monopolist,” but rather “positivist” or “realist.” It is both a public and private state, which solves the problem of order by taking the middle path and rejecting traditional extremes.

B. Political Theory vs. Public Policy

If the liberal theory of force sounds somewhat strange to most liberal philosophers, it will sound even stranger to public policy wonks. Until now, our public policy debates have not spoken of private prisons, police, and armies in the most fundamental terms. In lieu of first principles, they have offered the standard laundry list of . . . concerns: economic efficiency, egalitarian distribution, public goods, public accountability, human rights, industrial influence, market failure, and cultural commodification.

But . . . these standard paradigms are too broad to do the job well. . . . [B]readth is not an advantage because it masks the depth of the problem of order. . . . What is the most profound problem with privatizing the supply of military force? What . . . happens to the ultraminimal state? Does it perish at the hands of inefficient means? Maldistributions? Free riders? Accountability lapses? Human rights violations? Donations and bribes? Excessive fixed or sunk costs? A cultural faux pas?

. . . [E]ach of these suggests very good reasons that the liberal state should finance and allocate force. In addition, some identify reasons for the liberal state to provide and exercise force, as well—at least in some contexts, to some degrees. In the end, though, they all . . . highlight the general dangers of privatized institutions but not the special dangers of privatized force.

. . . [I]t is useful to revisit two public policy arguments to play out the contrast. First, the problem of order may sound like a problem of “agency costs,” but it is not—at least not in any conventional sense. It is true that the ultraminimal state wishes to make itself into a “principal” and the MO into an “agent.” But it cannot. Agency relationships are premised on the freedom of a market, which exists in the shadow of a dominant capitalist state. In the ultraminimal state, when the MO enters the market, its dominance destroys the market’s freedoms. The ultraminimal state is not free vis-à-vis the MO because there
is no state standing behind it, threatening to protect it. This is not just a failure of the
market; it is also a failure of the state itself in some sense. . . . The ultraminimal state fails
because it lacks a critical, cultural bond between . . . deliberation and domination.

In this respect, the liberal theory of force sounds distinctively cultural—much like the
“cultural commodification” critiques. There are similarities. The problem of order is a
problem about the social meaning of force. But it is not a problem of the popular meaning
of force, which is commonly called the problem of “legitimacy.” It is a problem of the
institutional meaning of force, which is why I have called it the problem of “loyalty.”

This distinction is critical, and it highlights the second difference between this theory and
others: Here institutions matter more than individuals. While the liberal theory of force
operates within the context of liberal states, and recognizes the values of democracy and
rights, it does not fetishize these values. It reminds us that in many liberal states, the
people are displeased with public institutions, but they do not revolt. It reminds us that
beneath our political and legal structures, there is a monopoly of force—a cultural bond
between the MO and the GO, which is much stronger than any generalized, popular
conception of legitimacy, democracy, or rights. Individuals may love the liberal state, but
generals, sergeants, and soldiers must love it. This military culture of loyalty is a matter
of the state’s political, legal, and economic survival.

C. The Problem of Order in Everyday Life

Since today’s liberal states are not ultraminimal states, it is tempting for today’s liberal
citizens to dismiss the problem of order out of hand. . . . [T]he United States has . . . an
extremely powerful, highly socialized network of military and policing organizations. But
we should not take this fact for granted. It is neither a necessary nor a natural fact. It
should be a source of awe, and perhaps fear, but it should not be an article of faith.

A brief tour of recent military coups and attempts should make us shift in our seats:
Venezuela (2002); the Philippines (2001); Fiji (2000); Pakistan (1999); Sierra Leone
(1996); the Soviet Union (1991); Thailand (1991); Nigeria (1990); Turkey (1980);
Argentina (1976); India (1975); Peru (1975); Portugal (1974); Greece (1973); Chile
(1973); and Brazil (1964).

“It can’t happen to us.” Perhaps. One certainly hopes not. . . . But . . . although the
ultraminimal state and the military state are fictions, the problem of order is utterly real.
It is the most vital problem of government. It is the only problem that the state cannot,
under any circumstances, resort to force to resolve.

Once we take the problem of order seriously, it becomes worthwhile to consider the
practical implications of the liberal theory of force. First, . . . [h]ow today’s “liberal”
states stack up against it? . . . [T]hey are doing fairly well. . . . [M]ost liberal states remain
stable and strong. Although private police are becoming more aggressive and expansive,
they have not seriously displaced the most aggressive and expansive functions of public
police forces yet. Although private armies are fighting wars, they are not leading us into
major wars yet. The liberal theory of force imposes limits, but for the most part, they are limits that we have not tested yet. Perhaps we never will.

. . . [But] there are already a few troubling signs. Because the liberal theory of force reads monopolization as socialization, it encourages us to focus on cultural connections between states and markets—for example, the fact that public police forces have started to model themselves on private policing corporations or that private military corporations have started training domestic and foreign troops. Are these examples of the private supply of force? Not exactly. But they undermine the cultural bonds between politicians, judges, and administrators on the one hand and officers and soldiers on the other. They undermine the “public” aspects of our public policing and military institutions.

D. The Ungovernability of Force

In closing, I must acknowledge some distinctly unpleasant—to some readers they might seem “illiberal”—aspects of the liberal theory of force. First, it is a positivist theory: It tends to presume the legitimacy of force rather than attempting to justify or explain it. Second, it is an institutional theory: It almost entirely ignores the role of individuals and focuses on relations between organizations. Third, it is a realist theory: It hinges upon the thesis that force is not ultimately governable—or, at least, that the most intense forms of force are the least governable.

In liberal terms, these might sound like surprising traits. At first glance, they seem to ignore or reject some of the most fundamental values of liberal thought. But the conflicts are largely apparent. The liberal theory of force neither ignores nor rejects these values. It recognizes them but highlights the profound constraints under which these values persist. What are these constraints? First: The concept of “legitimacy” is often deeply contested, so that “liberal” states must often govern in the absence of consensus. Second, institutional relations often overwhelm individual relations and limit our ability to vindicate individual rights. Third, and most importantly for our purposes: The most intense exercises of force often escape our best efforts at governance.

These constraints bring us back to that special, inherent tension between the concepts of liberalism and statehood that I explored in Part V. This tension is a decidedly familiar tension in liberal thought, which appears under many guises, in the oppositions between public and private, nationalism and liberalism, state and individual, autonomy and order, and freedom and force. In all such theories, it is there from the beginning, already embedded within the idea of the “liberal state.”

In the liberal theory of force, this tension emerges out of the collision between two old thoughts: First is an assumption about human nature—that individuals are rational and self-interested. Second is an assumption about the nature of force—that force corrupts. If these assumptions are not always correct, then we are fortunate folks. This, indeed, is the great aspiration of liberal government—the hope that, under the right conditions, power might not tend to corrupt us, or at least, that it might not succeed. Perhaps our “public” can master our power. It is a hope that puzzles me, but it is a hope that I share.
§ 9.2. Private Prisons


III. A Tale of Two Systems: Cost, Quality, and Accountability in Private Prisons

Private prisons are on the rise. Privately operated juvenile facilities—mostly community-based group homes or halfway houses—and federal adult halfway houses have been common in the United States since the 1960s. In 1979, private firms began contracting with the Immigration and Naturalization Service to detain illegal immigrants pending hearings or deportation. Private, large-scale investment in the construction and management of conventional prisons and jails dates from the mid-1980s. Prison privatization has been driven not only by the growing support among lawmakers and the public for private provision of traditional government services, but also by exploding prison populations resulting from stricter drug and immigration laws and changes in sentencing procedures.

By the end of 2000, there were 87,369 state and federal prisoners in private detention facilities in the United States—6.3% of all state and federal prisoners, and 22.7% more than in 1999. Of these, 15,524 were federal prisoners (10.7% of all federal prisoners) and 71,845 were state prisoners (5.8% of all state prisoners). The use of private facilities is concentrated in the South and the West. Texas and Oklahoma have the greatest numbers of inmates in private facilities; only six states—Alaska, Hawaii, Montana, New Mexico, Oklahoma, and Wisconsin, which combined account for approximately one-fifth of all state inmates—house over 20% of their prison population in private facilities. Privatization has been less widespread in local jails than in state prisons—only about 2% of jail beds are private—but jail privatization has been called the “next frontier” of privatization.

Comparative studies on the cost and quality of private and public prisons give reason to be cautiously pleased with private prison performance. The empirical evidence is consistent with economic theory, which predicts that with privatization, costs will fall and quality (however defined) may rise. The idealist could ascribe the satisfactory performance of private prisons to the power of market incentives; the cynic could point out that given public prisons’ bleak history and patchy present, private prisons perform satisfactorily compared to a rather low baseline. Each would be right.

Public prisons are not the most accountable of government systems; in fact, under certain circumstances, private prisons may be more accountable. In the qualified immunity context, recent Supreme Court decisions such as Richardson v. McKnight and Correctional Services Corp. v. Malesko have held private prisons to at least as high a

* Unsigned student note, part of the yearly student-written Developments in the Law.
standard of constitutional protection as public prisons. Judges’ and juries’ greater skepticism of private agencies than of government may also make private prisons more accountable; moreover, government oversight of private prisons may be less deferential than government oversight of its own operations. In addition, private prisons have substantially greater market accountability because they are concerned with winning new contracts and renewing old ones, and with avoiding both adverse publicity and drops in stock price. The continued promise of private prisons requires three concurrent innovations. First, evaluators must develop a rich set of performance measures, and prison data must be gathered and publicized. Second, the government must implement performance-based contracts that tie compensation to actual results. Finally, the government should maximize the efficiency gains from privatization and minimize opportunities for capture by institutionalizing competition between public correctional departments and private prison firms and making contract monitoring independent of both the public and the private sectors.

A. Private Prisons, Criminal Policy, and Democracy

Critics have argued that statutes authorizing private prisons unconstitutionally delegate core government functions to private parties. This contention has never been tested in court, but such arguments seem dubious given the uneven history of the nondelegation doctrine and the Supreme Court’s recent decision in American Trucking Ass’ns v. Whitman. True, private prison officials “determine when infractions occur, impose punishments and . . . make recommendations to parole boards,” but as long as they implement well-defined correctional policy with sufficient oversight, this delegation seems unobjectionable on federal constitutional grounds.

One modern-day objection to private prisons stems from opposition to corrections and criminal policy generally: if the problem is the incarceration of too many people, making prisons cheaper or more efficient is a false solution and may exacerbate the problem. Another objection is the “expressivist” critique “that to turn over responsibility for administering prisons and jails to private, for-profit companies at some level compromises the legitimacy of the state’s exercise of its authority to punish.” This Part takes a frankly consequentialist view of private prisons and this does not address these critiques. Private prisons might be inherently problematic under some moral theories and acceptable under others (both consequentialist and deontological), but space constraints preclude engaging this debate.

What about the specter of corruption? Industry lobbies government, and regulatory agencies can be captured by the entities they regulate; the private prison industry is no different. Not only may private prison companies lobby for preferential treatment, they may also, as entities that directly profit from incarceration, influence substantive criminal legislation by supporting tough-on-crime candidates, scaring the public about crime, and advocating tougher sentencing. The story is plausible, but it does not explain current levels of prison privatization or modern-day demand for more and cheaper prisons because the forces leading to the explosive growth of the prison population substantially predate the modern growth of the private prison industry.
Moreover, though private prison companies do lobby state and federal governments, so do prison guard unions, which also benefit from increased incarceration rates and prison construction. Prison guard unions generally contribute vastly more money to politicians than do private prison companies. The California prison guard union, for example, endorses and contributes millions of dollars to state candidates and “is among the largest campaign donors in the state.” Does privatization further distort criminal policy by replacing a single strong voice for incarceration with two voices? Or does the second, private, voice weaken the first by generally weakening the underlying public sector union? The answer is unclear.

Quite apart from whether political influence peddling distorts criminal policy, does such peddling weaken the case for privatization? Not necessarily, particularly when one considers different kinds of influence peddling: corruption and patronage. If a politician is corrupt and uses his power to extract money from the contractor, then privatization is likely to be inferior to public provision. Conversely, if a politician is involved in patronage and uses his power to pursue other political objectives, like serving politically powerful interest groups such as public employees’ unions, private provision is preferable.

B. Do Private Prisons Work?

1. Obstacles to Effective Assessment

Effectively evaluating prisons requires specifying goals and objectives, developing measures and indicators, and collecting comparison group data. Unfortunately, the political process does not value rigorous evaluation highly. Some states require only cost evaluation; only a handful require comparisons of both cost and quality, often to comply with statutory targets. Some neglect evaluation altogether. What monitoring data exist are often inadequate for outcome evaluation.

As if that were not enough, “cost” and “quality” are not clearly defined. Public agencies and private firms measure costs differently. Public prison budgets usually exclude various central administrative and support expenses, such as medical, legal, and personnel administration services, which other state agencies typically handle. Private budgets include these costs but do not include the government’s costs of preparing and monitoring contracts. As for quality, definitions differ across studies, and quality is difficult to compare in any case. If one facility has fewer assaults but more escapes than another, is it better or worse?

Few studies are rigorous. Even reasonably good studies leave much to be desired. No study seriously controls for many important factors that influence misconduct rates, such as staff-to-inmate ratios, custody technology, correctional policies, or age and race of inmates.

Furthermore, comparative studies do not adequately address serious overcrowding problems. Overcrowding, which may increase inmate violence and the incidence of infectious and stress-related diseases, thereby contributing to unconstitutional conditions,
is a serious problem in prisons and jails. Less expensive prisons allow for more capacity because the same prison budget can build more prisons, and increased capacity can relieve overcrowding. (Of course, cheaper prisons do not guarantee greater capacity, and greater capacity does not guarantee decreased crowding; still, it is reasonable to expect that cheaper prisons will not exacerbate crowding.) Thus, to the extent that private prisons decrease costs, privatization can improve conditions in both public and private prisons. Comparative studies between private and public prisons at a specific moment in time cannot register this across-the-board quality increase.

Finally, most studies do not analyze both cost and quality and thus are of limited value in assessing private prisons. Studies that do not look at both elements simultaneously cannot begin to analyze the costs and benefits of private prisons.

2. Evidence from the Studies

Studies that look at cost or quality alone do, however, provide some information. The most rigorous studies find clearly positive cost savings. On the quality side, comparisons are trickier, as there is no single metric representing quality. But none of the more rigorous studies finds quality at private prisons lower than quality at public prisons on average, and most find private prisons outscoring public prisons on most quality indicators. Most of these quality studies do not examine cost, but as private prisons are not expected to be more expensive, this result belies statements in the prison literature that assume that cost reductions must come at the expense of quality. Indeed, the few methodologically sound studies that evaluate both cost and quality suggest that cost is no higher in private facilities and quality is at least roughly equivalent.

Researchers at Louisiana State University compared three Louisiana prisons, one public and two private. The researchers concluded that the prisons were “as comparable as reasonably possible in terms of history, capacity, design, types of inmates, number, gender and ethnicity of inmates.” Privatization produced estimated cost savings of 12-14 percent (costs of $22.93 and $23.49 per inmate per day for the two private facilities, compared to $26.60 for the public facility). The quality comparison was a wash, with the private facilities faring better in some areas and worse in others. The private facilities, among other things, reported fewer critical incidents, provided safer work environments for employees and safer living environments for inmates, and had proportionately more inmates complete basic education, literacy, and vocational training courses. The public prison had fewer (zero) escapes and fewer aggravated sex offenses, more effectively controlled substance abuse among inmates, more consistently offered a broad education and vocational adult education program to inmates over four years, and provided more treatment, recreation, and rehabilitative services to inmates.

An Arizona study performed in 2000 by the state Department of Corrections compared three private prisons with fifteen public ones. The study found average savings of 13.6% at the private prisons in 1998 ($40.36 per inmate per day compared to $46.72 at public prisons) and 10.8% in 1999 ($40.88 compared to $45.85). The quality comparison gave the edge to neither group: private and public prisons had about equal inspection results, though public prisons generally had somewhat less severe disciplinary problems.
Finally, a 2000 study of Florida prisons compared the private South Bay Correctional Facility, operated by Wackenhut Corrections Corporation, with the most comparable public facility, Okeechobee Correctional Institution. After adjusting for various differences, including capacity, the study found that the private prison’s costs were 3.5% lower than the state’s costs for fiscal year 1997-98 and 10.5% lower for 1998-99, meeting the 7% cost reduction mandate established by law. Additionally, construction costs were 24% lower for the private prison. As for quality, the study found that South Bay had fully operational programs within six months of opening (as opposed to three years for Okeechobee), had fewer health service deficiencies than Okeechobee, was able to house more inmates three months after opening than Okeechobee could house after seventeen months, and implemented an innovative approach to housing certain “close management inmates.” On the negative side, two inmates escaped from South Bay in 1999, while none had ever escaped from Okeechobee.

3. That’s Fine in Practice, but How Does It Work in Theory?

Private prisons fare rather well in quality comparisons, but why? Contracts are necessarily incomplete: because the government and the private provider can only describe a general service and cannot specify beforehand in full detail exactly how the contractor should provide that service, the contractor has wide latitude in running the prison. This latitude permits the contractor to cut corners, reducing costs by reducing quality. It is no surprise that early economic models of privatization predicted that private ownership would reduce both cost and quality.

But more recently, economists have observed that cutting corners is not the only way to make money. It is easy to assume that an aversion to out-of-pocket costs will deter firms from implementing a “quality innovation”—but this assumption ignores opportunities for contract renegotiation. Because private prison companies can suggest such innovations to the government and renegotiate their contracts (or, in the real world, include extra services in a higher bid), they can capture some of the gains from quality innovation. They therefore have greater incentive to innovate in this way than their public counterparts, who cannot capture such gains.

Thus, while economic theory predicts that costs will decline under private management, it does not necessarily predict the same of quality. Whether quality increases or decreases overall depends on whether cost-cutting or quality innovation has a greater effect.

This simple model does not take into account other factors that may strengthen the case for privatization. For instance, the government may observe a provider’s harmful cost cutting and hold it against the provider in future rounds of bidding or decline to renew the contract. If even a few private providers compete against each other, the government can seek an alternative provider once it observes harmful cost cutting, rather than having to retake control of the prison. In addition, because of inefficiencies in current prison practices, there may be opportunities for cost cutting that does not reduce quality.
4. How Does the Private Sector Save Money?

The private sector saves money in several ways. First, private companies save money at the design and construction stage. They can typically design and build prisons in half the time required for governments to do so, because they can avoid the layers of red tape that play a role in safeguarding against government corruption but are arguably unnecessary when the government purchases a service from the private sector (using a procurement process that itself has safeguards against government arbitrariness). Private firms are also usually free of purchasing restrictions and subcontracting quotas. Contracting out prison design and construction reduces costs by 15 to 25 percent.

Second, private companies save money at the operation stage. The main savings come from reducing labor costs, both through lower wages and through more efficient use of labor. A private firm that had a role in designing a facility would be likely to use innovative design techniques that could minimize the number of guards required to monitor inmates. Moreover, because they are not bound by civil service rules in managing their personnel, private prisons use roughly one-third the administrative personnel of government prisons and use incentives to reduce sick time and consequent overtime expenditures. Private firms also save money by “maintaining tight control of inmates and keeping them well-fed and occupied with work, education, or recreation.” Finally, private firms are free from many bureaucratic purchasing rules and can often buy supplies at lower cost than the government.

C. Public and Private Accountability

Abuses happen in any system. But how do different systems react to abuse when it occurs? While there is no systematic information about the reaction to prisoner abuse in public and private prisons, case studies may provide a flavor of the accountability mechanisms at work.

1. Private Prisons’ Legal Accountability

The traditional hostility of juries toward corporate defendants, private prison guards’ inability to claim qualified immunity in § 1983 civil rights suits, and courts’ unwillingness to defer to the judgment of corporations all increase private prisons’ legal accountability relative to public prisons.

a. Jury Hostility

Empirical studies have found that juries are more likely to award large verdicts against corporations than against governments. Jury hostility also affects settlements, which are made in the shadow of expected recovery amounts at trial. Consider, for example, the case of the Northeast Ohio Correctional Center, a federal prison in Youngstown, Ohio, run by CCA under a contract with the District of Columbia. As part of a recent settlement of a lawsuit alleging inadequate security and medical treatment as well as excessive force, CCA paid $1,650,000 in damages to the 2000 members of the inmate class—an extraordinarily high settlement amount for class actions involving prisoners. This huge settlement amount probably reflects the well-known tendency of juries, rightly or
wrongly, to be less sympathetic to large corporate defendants. Moreover, monetary awards against public prisons are more limited.

b. Qualified Immunity: The New Bifurcated Regime

A danger inherent in privatization is that public responsibilities will be performed by private individuals without effective oversight. Recent cases, however, have sought to avoid this pitfall and, in some ways, hold private prisons to an even higher standard than public prisons. Under *Harlow v. Fitzgerald*, government officials (including prison guards) performing discretionary functions are shielded from liability for civil damages if their conduct does not violate “clearly established statutory or constitutional rights of which a reasonable person would have known.” In *Richardson v. McKnight*, however, the Supreme Court held that private prison guards do not enjoy this qualified immunity. Thus, *McKnight* bifurcates the treatment of public and private prisons in a way that makes private prisons more, not less, accountable.

Does this dichotomy make sense? If governments are cost-minimizers, it does not. In the presence of liability, cost-conscious governments would hire timid guards, who require less indemnification, for public prisons, and would choose timid contractors, who submit a lower bid, for private prisons; thus, the argument for qualified immunity would not depend on whether the prison is public or private. But governments often do not minimize costs. They often face soft budget constraints, may award contracts to friends or political allies, and may monitor prisons more or less depending on their political ideology. Because government officials’ motivations are complex, it is difficult to predict how the effects of qualified immunity will vary as between public and private prison operators. The efficiency arguments for only extending qualified immunity to public guards are thus somewhat indeterminate.

Considerations of compensation and accountability, however, cut against immunity for both public and private prisons. On the compensation side, immunity presumably makes a difference in some cases, and where it does, it may prevent victims from being compensated. On the accountability side, immunity allows prisons to externalize the cost of constitutional violations onto prisoners—not a particularly well-represented segment of society.

Nor does qualified immunity seem necessary to reduce the drain on the court system. Most § 1983 prison lawsuits are considered frivolous, and almost all fail before getting to the merits. Qualified immunity would probably not change inmates’ filing behavior, and even if it did, it is not clear that this would be desirable, because our system deliberately tolerates a certain amount of meritless litigation to avoid missing the occasional worthwhile claim. A regime of liability, with high procedural hurdles (such as requirements of exhaustion of administrative remedies) or penalties for frivolous litigation (such as the frequent-filer penalties in the Prison Litigation Reform Act), may better serve the goals of compensation and accountability for both government and private actors.
Thus, given the bifurcated regime established in *McKnight*, private prisons are, if anything, more accountable for their constitutional violations than are public prisons. The presence of this additional judicial check should in turn increase private prison quality.

**c. Other Legal Accountability Mechanisms**

Not only may privatization make juries more skeptical of prisons and guards more susceptible to suit, but it may also make judges less deferential. Courts have traditionally been deferential to the government in prison suits because of the common judicial unwillingness to second-guess the political branches. Though this unwillingness is understandable, deference is the enemy of robust accountability. Yet courts have often been hostile to private, for-profit delegations; while the nondelegation doctrine is almost never applied, judicial hostility manifests itself through more vigorous due process scrutiny. As administrative power moves to private prison corporations, courts may abandon their deference to prison officials, because conflicts of interest will appear more obvious and contractual terms will necessarily be incomplete.

In addition to litigation, government monitoring adds a further layer of review. While public prisons have government monitors, monitors of private prisons (even if captured to some degree) are likely to be more independent than monitors of government-run prisons.

**2. Private Prisons’ Market Accountability**

Market mechanisms, such as governments’ ability to rescind or decline to renew private firms’ contracts, and more generally, the potential for bad publicity to cause a drop in firms’ stock prices, further increase private prison companies’ accountability.

**a. Contract Rescission**

A private prison may have its contract rescinded. This possibility is not always as easy as it sounds—the private prison industry is a somewhat concentrated oligopoly, though that may change with increased privatization. But as long as more than one firm is operating and the government continues to run part of the prison system, someone will be available to take over a dysfunctional prison, making the government’s threat to rescind a contract somewhat credible. At any rate, even such imperfect discipline is more difficult to impose on public prisons—the government cannot take over its own prison except by firing civil servants, and it cannot have a private firm take it over except by opening a new bidding process, which is more difficult than finding someone to take over an existing contract.

In 1995, for example, an investigative reporter and a surprise inspection revealed appallingly crowded, unsanitary, and abusive conditions at the Bowie County Correctional Facility, a private “warehouse turned prison facility” in Texarkana, Texas, housing 500 inmates from Colorado. After a federal lawsuit and a riot, the state of Colorado terminated Bowie County’s contract and relocated its inmates.

At the Seal Beach City Jail, operated by the private Correctional Systems, Inc. since 1994, two private prison guards were indicted in August 2001 for allegedly “arranging
and concealing an attack on a drunken inmate who was singing boisterously in the jail’s detoxification cell.” Within ten weeks of the attack, both guards had been fired; one guard was charged with federal civil rights violations and the other with being an accessory after the fact for trying to conceal the attack. The first guard was sentenced to over four years in prison. This incident prompted the city to review its contract with CSI and to consider other options.

Similarly, in Santa Fe County, New Mexico, CCA brought in prisoners convicted of murder and rape from Oregon to fill its private jail cells but reportedly failed to notify the Sheriff’s Department that it was housing inmates who posed a danger to the public. The Oregon prisoners were removed from the jail only after county officials threatened to cancel the county’s contract with the company.

b. Adverse Stock Price Effects

Private corporations are sensitive to drops in their stock prices. Contract rescission, as well as the possibility of lawsuits with high damage awards, affects profitability, and perceptions of a company’s profitability are reflected in the price of its stock. Thus, a private corporation is punished financially for bad news, and possibly for mismanagement that may impose costs in the future. For example, the INS detention center in Elizabeth, New Jersey, run by Esmor Corrections, erupted in a massive riot in 1994 because the company had continuously cut corners on food and facility upkeep. Esmor’s stock price dropped from $20 a share to $7 after news of the riot became known.

While fear of stock price drops may make private prisons conceal their problems—CCA has been accused of covering up escapes from the Youngstown prison described above—such secrecy has its limits, because escapes, riots, and prisoner litigation are hard to cover up. Following disclosure of the problems at the Youngstown prison, CCA’s stock hit a one-year low. A more plausible story is that private companies are more concerned with keeping their stock prices high over the long term by insisting on sound management and that guards and wardens can be encouraged to act responsibly through stock ownership in the company.

c. Greater Market Flexibility

Not only are private prisons under greater market pressure to keep conditions from getting out of hand, they are also more able to change because they are free from some of the constraints of government management. In July 2001, for instance, four hundred prisoners in a CCA prison in Kentucky started a riot in which “inmates set[ed] mattresses on fire and toss[ed] TVs and toilets through the windows. Two weeks later, CCA fired the warden and his top assistant, citing ‘policy violations.’” Government employment protection, in this case, would have been an obstacle to disciplining the offending prison officials.
3. Are Public Prisons Any More Accountable Than Private Prisons?

a. The Dynamics of Organizational Stasis

Susan Sturm describes the “dynamics of organizational stasis,” which “lock in the current conditions in prisons” (both public and private) and “disable regular prison participants from achieving institutional self-correction”:

Even in the face of riots, violence, or public exposure of brutal conditions, legislatures frequently provide minimal support for change. The familiar study commission or task force conducts a public hearing or investigation, culminating in a report with recommendations that are infrequently enacted into law. Even when the legislature does respond to abuse with legislation, there is little accountability for its enforcement and the administration may comfortably ignore it without legislative sanction.

Much of Sturm’s description of prisons applies equally to public and private prisons—inmates are generally powerless to change prison conditions, and both guards and administrators often resist reform. Other elements, however, apply primarily to public prisons: unions concentrate on preserving their members’ pay, benefits, and seniority, and the budgetary process, in which this year’s allocation is presumed to be the baseline for next year’s, supports the status quo. It is well known that public bureaucracies have different incentives and attitudes toward change than do private companies. Companies concerned about winning a bid, retaining their contracts, maintaining their stock price, or just being marginally more protected against prisoner lawsuits seem more likely to overcome their institutional stasis.

b. Corcoran State Prison

Contrast the above factors, which promote accountability in private prisons, with the sad tale of Corcoran State Prison, a California public prison, in which guards killed seven inmates and wounded forty-three others between 1989 and 1995. Rival gang members fought in human cockfights overseen by prison guards; officers abused and beat inmates; problem inmates were disciplined by being locked in a cell and then beaten or raped by an inmate enforcer dubbed the “Booty Bandit.”

Conditions at Corcoran are not typical of public prisons. What is striking about Corcoran, however, is how little was done to alter these conditions even once they were discovered. The Corcoran episode serves as a reminder of the weakness of public prisons’ accountability. When replacing one’s service provider is difficult—and especially when the provider is identified with the purchaser, capturing the policy advice process—moral outrage at abuses can only get one so far.

In 1994, the FBI began its probe of Corcoran, and in 1996, prompted by media attention, the state launched a pair of investigations, one by the Department of Corrections and another by the Attorney General’s office. These investigations were stymied by political pressure from the governor’s office and the prison guard union, and the state probes yielded not a single criminal charge. Instead, state investigators “spent considerable
manpower trying to dig up dirt” on the whistleblowers who had reported these abuses to the FBI. One whistleblower, Richard Caruso, was the only officer disciplined as a result of the state investigations (though he sued and received a large settlement from the state, the largest amount ever given to a whistleblowing officer in California). Another whistleblower changed jobs within the prison system because of alleged retaliation by prison officials (and recovered nothing). The FBI inquiry ultimately resulted in charges against eight officers, but all were acquitted in June 2000.

**D. The New Frontier: Information, Contracts, Monitoring**

While private prisons today provide acceptable quality at a lower cost than do public prisons, they will only continue to do so as long as their buyers—federal and state governments—care. Bad things still happen at private prisons. Private prison companies have not always been forthcoming with information about the types of inmates they are holding, and reports of maximum-security prisoners being housed in portions of private prisons designed for the general population are not uncommon. Effective judicial, legislative, and administrative oversight continues to be necessary. Perhaps the scope of prison litigation should be expanded to keep both public and private prisons honest, but there is no substitute for performance contracts that encourage quality improvements, effective monitoring, and information gathering and disclosure.

**1. Performance Measures**

Susan Sturm notes that prison officials often know little about their own operations. Prison administrators gather some information to prove the prison’s compliance with court orders, but it would be preferable for them to do so without being forced. One way to increase private prisons’ disclosure, at least in the case of federal institutions, is by subjecting them to the same Freedom of Information Act disclosure requirements as public prisons—though this, too, is a piecemeal solution.

Perhaps the environmental information regime can provide a model for the disclosure of prison information. The Toxics Release Inventory (TRI), which requires that industrial facilities report the release and transfer of specific chemicals, has been credited with dramatic reductions in pollution. In part because of the existence of these reporting requirements, much formerly fragmented and incomparable data has become standardized and can be aggregated to allow comparisons of different firms or states, or to track performance over time.

Of course, TRI mandates only that information be gathered and made public, not that anything be done with that information. But the mere availability of information, whether environmental information about industrial facilities or correctional information about prisons (floor space, number of violent incidents, recidivism), has real effects. First, “you manage what you measure”—administrators have a natural desire to improve what they have data about. Second, information gathering encourages peer monitoring within an industry, and industry self-regulation can be a valuable supplement to other forms of monitoring or control. Third, information is valuable to regulators, would-be regulators, industry actors seeking to stave off regulation, community monitors and “informal
regulators,” capital markets, labor markets, and customers (in the prison case, state and federal government entities).

Information as regulation has its critics. Regulation proponents charge that information does not guarantee any action or results and that it is most useful in combination with other regulation. From the other side, critics charge that because information is highly contextual, any mandated information will be too simplistic and perhaps misleading, especially when presented at a high level of aggregation. But while particular information initiatives may be flawed, and while mere disclosure may not be sufficient, collecting and publicizing a rich (and meaningful) set of performance indicators will increase the incentives for improvement that are already inherent in the competitive contracting process.

2. Improving Performance Through Contract

At best, contracts “represent potentially useful accountability instruments [and] vehicles for achieving public law values, such as fairness, openness, and accountability.” Private prison contracts should have specific terms, graduated penalties, and strong oversight by a “contract manager” working for the public agency; they should require private prisons to “observe minimal administrative procedures such as notice and hearing requirements,” and perhaps explicitly give inmates or surrounding communities third-party beneficiary rights, which would allow oversight through contract litigation when government oversight fails. States could also require, as contractual terms, compliance with American Correctional Association and National Commission of Correctional Health Care standards.

Moreover, states could mandate that private prisons provide the same training that is required of public prison guards, though requiring certain inputs is presumptively less effective than looking to outcomes where these are measurable. Contracts could require that private firms carry civil rights liability and other insurance; that they disclose conflicts of interest; that they allow access to records and entry to the facility by inspectors; or that they be independently monitored or audited by certified professionals. Finally, contracts could tailor termination rights and provide for easy amendment of contractual terms. In short, contract designers can be highly creative—and thorough—in writing accountability into contractual terms.

Most basically, corrections departments should move toward performance-based contracts. Ideally, performance-based contracts should “clearly spell out the desired end result” but leave the choice of method to the contractor, who should have “as much freedom as possible in figuring out how to best meet government’s performance objective.” These contracts also structure contractor payments to encourage the desired results, rewarding the contractor for improvements and penalizing it for poor performance or rising costs.

This approach seems feasible for corrections. The American Correctional Association is revising its accreditation standards to include performance measures, and the Office of Juvenile Justice and Delinquency Prevention is developing performance-based standards.
for juvenile correctional facilities. Performance measures for prisons could include process measures such as the number of educational or vocational programs, or outcome measures such as the Logan quality of confinement index, the number of assaults, or the recidivism rate. Governments could even require that contractors pay for elements that are often externalized, such as the cost of escapes. Because no single statistic adequately captures “quality,” and because focusing on any single measure could have perverse effects, performance-based contracts should tie compensation to a large and rich set of variables.

3. Awarding and Monitoring Contracts To Maximize the Gains from Privatization and Minimize Capture

Accountability can also be improved by redesigning correctional agencies. Under the basic model of accountability, the public correctional agency sets contract terms, awards contracts, and monitors compliance. Capture is frequent, cross-fertilization rare.

But there are other models of accountability. In the United Kingdom, “the role of the chief inspector of prisons brings some external scrutiny to the prison system generally, including the private sector . . . . This aspect of the accountability system is vigorously independent, with no danger of capture to date.” However, the chief inspector’s reports are not binding, and the government has resisted attempts to give the office more teeth. In Florida, private prisons have their own regulatory authority that operates independently of that of public prisons; the private prison monitor is in turn independent of the private prison authority. This system minimizes capture, but it also minimizes cross-fertilization. Moreover, “[t]here is also a danger that the monitors may develop a loyalty to the Privatization Commission which in turn might inhibit their willingness to make public criticisms—a variant of the capture principle.”

Perhaps the most promising model is one suggested by British penologist Richard Harding, in which both public and private prisons are subject to, and monitored by, an independent body that takes over new prison projects from the outset and allows both public and private systems to bid on them. After the contract term expires, both public and private systems may bid on the prison again, so that facilities can move among private firms and between the private and public sectors, promoting both accountability and cross-fertilization. This model separates the roles of government as purchaser and provider, making it more difficult for service departments to capture the policy advice process and use their power to recommend themselves as service providers.

E. Conclusion

This Part has not engaged the “softer” arguments about privatization, but has rather taken correctional policy as given and prison privatization as ethically neutral in itself. Does political influence peddling weaken or strengthen the case for privatization? It depends whether corruption by corporations is worse than patronage of public prison guard unions—a question that calls for further research.
Turning to the cost and quality comparisons, what imperfect empirical evidence there is suggests that private prisons cost less than public prisons and that their quality is no worse; it is perhaps unsurprising that prison privatization behaves in this respect much like privatization of other state and local services. Moreover, there are many reasons to believe that private prisons are more accountable than public prisons—both because of heightened legal and market accountability for private forms and because accountability in the public sector is so limited. There are also numerous untapped opportunities for improving private prisons even further with richer information gathering and dissemination, performance contracts, and better monitoring. In short, despite all of their possible faults, private prisons are a promising avenue for the future development of the prison system.


To date, the debate over private prisons has focused largely on the relative efficiency of private prisons as compared to their publicly-run counterparts, and has assumed that, if private contractors can run the prisons for less money than the state without a drop in quality, then states should be willing to privatize. This “comparative efficiency” approach, however, has two significant problems. First, it is concerned exclusively with efficiency, despite the fact that the privatization of prisons arguably implicates more urgent values. Second, it accepts the current state of public prisons as an unproblematic baseline, thus failing to consider the possibility that neither public prisons as presently constituted nor private prisons in the form currently on offer are adequate to satisfy society’s obligations to those it incarcerates. In this Article, Professor Dolovich examines the private prisons issue from a third perspective, that of liberal legitimacy. On this standard, if penal policies and practices are to be legitimate, they must be consistent with two basic principles: the humanity principle, which obliges the state to avoid imposing punishments that are gratuitously inhumane; and the parsimony principle, which obliges the state to avoid imposing punishments of incarceration that are gratuitously long. After sketching the foundation for this legitimacy standard, Professor Dolovich then applies it to the case of private prisons. Approaching the issue of private prisons from this perspective helps to reframe the debate in two ways, both long overdue. First, it allows for a direct focus on the structure and functioning of private prisons, without being derailed by premature demands for comparison with public-sector prisons. It thus becomes possible to assess directly the oft-heard claim that the profit incentive motivating prison contractors will distort the decisions made by private prison administrators and lead to abuses. Second, it makes it possible to see that the state’s use of private prisons is the logical extension of policies and practices that are already standard features of the penal system in general, thus throwing into sharper relief several problematic aspects of this system that are currently taken for granted. In this sense, the study of private prisons operates as a “miner’s
INTRODUCTION

During the 1980s and 1990s, the population of America’s prisons and jails soared to unprecedented levels. Watching the cost of incarceration rise accordingly and finding themselves responsible for many more inmates than they were able to accommodate in existing facilities, state officials turned to the private sector for help. They were met by entrepreneurs offering a range of services designed to appeal to the overtaxed prison administrator, including everything from the siting and building of new prisons to the day-to-day management of whole inmate populations. By 2003, over 90,000 inmates across the country were housed in prisons and jails run by for-profit prison-management companies.

This emergence of privately run, for-profit prisons, or “private prisons,” sparked a heated debate, at the heart of which has been one basic question: should responsibility for offenders convicted by the state be delegated to private, for-profit contractors, or should incarceration continue to be administered exclusively by public institutions staffed by state employees? The private prisons issue has thus widely been viewed as a choice—even a competition—between alternative organizational forms.

For the most part, debate on this issue has focused on the relative efficiency of private prisons as compared to their publicly run counterparts and has assumed that, if private contractors can run the prisons for less money than the state without a drop in quality, then states should be willing to privatize. There are, however, at least two significant problems with this “comparative efficiency” approach. First, it is exclusively concerned with the value of efficiency. Such a focus may be appropriate in many contexts in which privatization is contemplated, but it is not so in the prison context. Incarceration is among the most severe and intrusive manifestations of power the state exercises against its own citizens. When the state incarcerates, it strips offenders of their liberty and dignity and consigns them for extended periods to conditions of severe regimentation and physical vulnerability. Before seeking to ensure efficient incarceration, therefore, it must first be determined if the particular penal practice at issue is even legitimate.

Second, in its drive to assess the relative performance of private prisons, comparative efficiency accepts the current state of public prison conditions as an unproblematic baseline. Comparative efficiency asks: how do private prisons compare with their public-sector counterparts? And in terms of conditions of confinement, this standard is satisfied when conditions in private prisons are shown to be as good as conditions in the public prisons they seek to replace. Whether the baseline set by public prisons is itself good enough to meet any justifiable objective standard is never considered. The conversation as defined by comparative efficiency is thus framed to sidestep, rather than directly engage, the fact that conditions in many prisons—public and private alike—fall far short of satisfying society’s obligations to those it incarcerates.
Not all participants in the private prisons debate take comparative efficiency to be the appropriate standard. For a small group of critics, what matters most is not the relative efficiency of private prisons but their perceived lack of legitimacy. These critics share the view that incarceration is an inherently public function and thus that recourse to private prisons is inappropriate regardless of the relative efficiency of this penal form.

By introducing into the debate a concern for the legitimacy of private prisons, this “inherent public function” approach makes a laudable attempt to shift the inquiry beyond the bounds of comparative efficiency and into the terrain of the more explicitly normative. Yet this alternative framework also has its shortcomings. For one, its dismissal of the relevance of private prisons’ practical implications for the prisoners themselves seems both coldhearted and blind to the significance of the humanity of actual sentences served for the legitimacy of a given punishment. Moreover, although its governing standard—legitimacy—is more appropriate for the prison context than efficiency, the inherent-public-function approach also accepts the status quo as an unproblematic baseline for analysis, and thus is ultimately trapped in the same unduly narrow frame as comparative efficiency. On both approaches, the only relevant question is the comparative one—that of whether states should keep operating the prisons themselves or be willing to turn this task over to private, for-profit contractors. Neither approach, therefore, is able to consider the possibility that neither public prisons as presently constituted nor private prisons in the form currently on offer represents an acceptable choice.

In this Article, I approach the private prisons issue from a third perspective, that of liberal legitimacy. Liberal legitimacy offers what the private prisons debate has thus far lacked: an independent normative standard for assessing the legitimacy of penal policies and practices in a liberal democracy. On this standard, if our penal policies and practices are to be legitimate, they must be consistent with two basic principles: the humanity principle, which obliges the state to avoid imposing punishments that are gratuitously inhumane; and the parsimony principle, which obliges the state to avoid imposing punishments of incarceration that are gratuitously long. In each case, gratuitous punishment is that which cannot be justified to all members of society under fair deliberative conditions. In this Article, I sketch the foundation for this legitimacy standard. I then apply its conditions to the case of private prisons. Doing so reveals the extent to which punishments served in private prisons fall short of society’s obligations to those it incarcerates—and why they do so.

Liberal legitimacy thus rejects the comparativist’s impulse that has thus far defined the private prisons debate. The question here is not: how do private prisons compare with public prisons? It is instead: to what extent is the use of for-profit private prisons consistent with society’s obligations to those it incarcerates?

Asking this latter question helps to reframe the debate in two ways, both long overdue. First, it allows for a direct focus on the structure and functioning of private prisons, and thus for the development of a rich understanding of how the private prison system actually works and precisely where it fails, without being derailed by premature demands for comparison with public-sector prisons. Most notably, it allows for direct assessment
of the claim that the profit incentive motivating prison contractors will distort the
decisions made by private prison administrators and lead to abuses. This claim is often
raised by opponents of private prisons. Yet it is rarely pursued in any sustained way, for
whenever it is voiced, it is either dismissed outright as unsupported or quickly deflected
by reference to the admittedly incontrovertible fact that public prisons, too, are rife with
abuse. By contrast, assessing private prisons against the standard of liberal legitimacy not
only allows but demands a thorough analysis of the concern that prison contractors’ profit
motive will lead to cutting corners in ways likely to harm inmates. It thus enables an
understanding of the dangers posed by private prisons that is at once more comprehensive
and more nuanced than is possible from within the comparativist frame.

Second, confronting the ways in which private prisons are at odds with society’s
obligations to those it incarcerates provides the basis for a far-reaching critique of several
practices that currently inform prison administration more broadly. The possibility that
studying private prisons might afford a fresh perspective on society’s penal practices in
general has not been seriously considered by those engaged in the private prisons debate.
In fact, a guiding premise of this debate has been that for-profit private prisons represent
a radical departure from the way the public prison system otherwise operates. But this
premise is false. Although private prisons do have some distinctive features, the
differences between public and private prisons are mostly differences of degree. The use
of private prisons is thus neither an isolated nor an aberrant approach to punishment, but
is rather the logical extension of policies and practices that are already standard features
of our prison system. Examining private prisons from the perspective of liberal
legitimacy exposes this overlap, thereby throwing into sharper relief several problematic
aspects of the penal system as a whole that are currently taken for granted. In this sense,
the study of private prisons operates as a “miner’s canary,” warning that not just the
structure of private prisons, but also that of American punishment practices more broadly,
may need reconsideration.

This Article proceeds as follows. Part I offers a brief account of the history and
reemergence in the late twentieth century of private sector involvement in American
corrections. Part II addresses the question of how to translate a commitment to the
baseline liberal democratic values into the basis for public-policy critique. In doing so, it
presents the framework of liberal legitimacy to be applied to private prisons. Part III
considers the use of private prisons from the perspective of the humanity principle. In
particular, it examines the incentives prison contractors face to reduce costs in ways
likely to cause harm to inmates, and argues that existing mechanisms for checking
contractors’ tendencies in this direction are inadequate to the task. Drawing on this
analysis, Part III identifies two practices of prison administration that threaten the health
and safety of prisoners: (1) the contracting out to for-profit entities for the provision of
essential prison services to save states money on the cost of corrections, and (2) the
delegation to prison officials of considerable power and discretion over prisons and
prison conditions absent adequate accountability mechanisms. Although, as Part III
shows, the dangers these practices create are heightened in private prisons—a fact that
explains the elevated levels of violence in private prisons as compared with public
prisons—these practices are standard fare in public prisons as well. Part III thus
concludes by identifying steps prison officials ought to take to curtail these practices or
mitigate their harmful effects, even where the prisons themselves are run by public employees. Finally, Part IV examines private prisons from the perspective of the parsimony principle. In contrast to Part III, Part IV introduces a set of considerations that has not previously been addressed as such in the literature. It is thus necessarily more speculative and suggestive than Part III. Still, Part IV does identify a further practice that threatens the legitimacy of punishment in general: political advocacy in the sentencing-policy arena by actors with a strong financial interest in increased incarceration rates and longer prison sentences. As Part IV suggests, this practice creates reason for concern whether the advocacy groups in question are private prison providers, correctional officers’ unions, or voters in rural districts who view prison building as a possible source of economic development.

I. THE EMERGENCE OF THE MODERN PRIVATE PRISON

A. Historical Antecedents

The involvement of private interests in American corrections began long before the current generation of private prison companies emerged—indeed, even before the existence of the prison as we know it. Before addressing the modern private prison, it is worth briefly considering some of this earlier history, which raises themes that will inform later analysis.

In colonial America, the meting out of criminal punishment was purely a local matter and could include any of a range of sanctions, among them fines, flogging, the stockade, banishment, and the gallows—but not imprisonment. As in eighteenth-century England, jails were merely holding chambers for debtors or for those individuals awaiting trial or punishment. Jailors paid for the running of the jails themselves and were reimbursed by the county according to a fee schedule. They also routinely supplemented their income by taking bribes from prisoners in exchange for certain privileges and charging prisoners for meals and alcohol. The less money spent on upkeep, the more money the jailor made; jails were thus generally overcrowded and extremely unsanitary.

It was in the late eighteenth century that criminal punishment in America came to take the form of incarceration for a set period in a penal facility. In the early penitentiaries, prison labor was introduced as part of rehabilitative programs, but it quickly became the means through which state governments could recoup the costs to the state treasury of imprisoning criminals. Indeed, the history of nineteenth-century American prisons is a history of contracting between the state and private interests for the use of convict labor in efforts on both sides to achieve financial gain. These contracts took many forms. In some cases, as with New York’s Auburn penitentiary, contractors would supply the raw material and collect the finished product at the end, with the work taking place at the prison. In others, as in Louisiana, the state leased its entire penitentiary to a private contractor, who then assumed the cost of running the facility in exchange for the labor of its inmates. The most common arrangements, however, involved the leasing of convict labor for work on plantations, on railroads, in mines, or in other labor-intensive industries.
Although convict leasing was found throughout the nineteenth-century United States, it was most widely used in the Southern states after the Civil War. This development was in part a function of the serious financial straits of the former Confederate states in the postwar years; convict leasing offered a way both to defray the costs of incarceration and to rebuild the shattered Southern economy. At the war’s end, demand for convict labor was high, as those who had previously relied on slave labor found themselves in need of a pool of cheap workers. Both in response to this demand and as a way for white society to reassert its power over the newly emancipated black population, the Southern states began to increase dramatically the sentences exacted against petty criminals, the vast majority of whom were former slaves. For example, in 1876, the Mississippi legislature passed a “major crime bill,” known as the “Pig Law,” which redefined the crime of grand larceny to include “the theft of a farm animal or any property valued at ten dollars or more.” Violation of this law, which was “aimed directly” at the newly freed slaves, meant up to five years in state prison. Moves like this one accompanied the legalization of convict leasing and ensured sufficient convicts to meet the demand.

Convict leasing uniformly meant the severe abuse of leased convicts, thoroughly inadequate living conditions, and utter indifference as to whether they lived or died. Because the prisons ensured a steady supply of convicts, from the contractors’ perspective one convict was as good as another. Many contractors therefore routinely worked their charges literally to death.

Historical accounts of inmate labor contracts in nineteenth-century America reveal that the practice was plagued by more than inmate abuse. In addition, state after state found itself being outmaneuvered and taken advantage of by the private parties with whom the state had contracted for the labor of its convicts. In California, for example, the state tried in 1858 to rescind a contract for the labor of inmates at San Quentin when it became known that the contractor, John McCauley, had “blatantly violat[ed]” the terms of the contract “to squeeze as much out of the arrangement as possible.” McCauley had “ignored the physical needs of the convicts, ignored the orders sent down from Sacramento, ignored the suggestions of his own prison officers, ignored everything but his profit.” McCauley fought the rescission in court and won, and the state, which had entered the contract in the first place to save money on the running of San Quentin, had to pay over $200,000 to buy him out.

The predominant theme of accounts of prison labor contracts gone awry is the state’s vulnerability to nonperformance by its contracting partner once the state had divested itself of responsibility for its prisoners. In Virginia, Nebraska, and Tennessee, the story was the same: the state leased its convicts to private interests, discovered violations of the contract terms directed to increasing the profit of the contractor, and found itself unable to cancel the contract. The reasons for this incapacity varied from state to state and included the lessee’s political connections (as in Louisiana and Kentucky), the state’s dependence on the contractor to provide for the prisoners’ needs (as in New York, where in 1851 the wardens of Auburn penitentiary were forced to give significant concessions to the contractor running an on-site carpet shop or leave “idle more than 300 inmates” and risk the loss of necessary revenue), and the risk that courts would side with the contractors (as in California), thus forcing the state to pay dearly to regain state control of
its prisons. In each case, for these various reasons, once the contracts had been signed, the balance of power shifted to the contractors.

It would be a mistake to draw too many conclusions from this history for the current chapter of private sector involvement in prisons. The contemporary experience is governed by a set of norms, not in place a century ago, forbidding the economic exploitation and physical abuse of inmates. Today, there is also a stricter standard of political accountability, an extensive public bureaucracy with the capacity to regulate and administer complex institutions, and the default expectation that the state bears the burden of financing the prison system. But as will be seen, this history does introduce certain themes arising from private involvement in corrections that are still relevant today.

B. The Corrections Crisis of 1980s America

The reemergence of private contractors in American corrections is traceable to the dramatic growth in incarceration nationwide over the past three decades. In 1985, there were over 740,000 people behind bars, up from 226,000 ten years previously. By 1990, this number had hit 1.1 million, by 1995, it was almost 1.6 million, and by 2003 it was over 2.1 million. For legislators and prison officials around the country, this incarceration explosion created some vexing practical problems: Where to put all the convicted offenders? And how to pay the bills?

Initially, state officials nationwide responded to the first of these problems—finding room for all the bodies—by shipping convicted offenders to existing penal facilities and letting the wardens sort it out themselves. The limitations of this approach, however, were soon clear, as prisons and jails quickly came to be operating well over capacity. Eventually, the courts began issuing orders requiring government officials to relieve the overcrowding, and it became apparent that more prisons had to be built. But this solution, too, had its problems: building prisons is not cheap, and the building process itself can be complicated and time consuming. Jurisdictions urgently needing new prisons thus faced—and still face—serious obstacles to getting new facilities up and running.

Nor was the second practical problem caused by the rapidly increasing incarceration rate—how to foot the bill—fully contained in the cost of building new prisons or renovating old ones. Prisons also have operating expenses, and these costs, too, can be high. Corrections officials have to be trained and their salaries and benefits paid; inmates’ food, clothing, medical care, programming, security, and so on, must be provided for; overhead must be covered. In all, by the mid-1980s, many states were facing serious budgetary problems traceable to the increased cost of running their prison systems, and these problems have only grown in intensity as incarceration rates have continued their upward climb.

C. Enter the Private Sector

It was under these circumstances that the states turned to the private sector for help. The help offered took two forms. First, the private sector offered to assist states with the
capital financing of prison construction, a version of private sector involvement known as "nominal privatization." Second, private firms offered to take over the day-to-day management of entire penal facilities, pledging to run the prisons at a lower cost than the state would otherwise pay. At the time these firms emerged, this latter form of privatization—"operational privatization"—was not a new idea; in the late 1970s, the federal Immigration and Naturalization Service (INS) had begun contracting with private firms for the building and operation of holding facilities for illegal immigrants awaiting hearings or deportation. What was new was the status of those to be housed in the privately managed facilities—adults convicted of crimes and sentenced to state custody as punishment. Notwithstanding this difference, the process of privatizing the prisons took the same form as with the previously privatized INS facilities: states deciding which facilities to privatize and issuing a request for proposals (RFP), firms bidding for the contract, and the winning firm getting a set payment per inmate per day in exchange for assuming responsibility for running the facility and providing for inmates’ needs. There are variations on this standard arrangement, but the basic idea in each case is the same: “the state remains the ultimate paymaster and the opportunity for private profit is found only in the ability of the contractor to deliver the agreed services at a cost below the negotiated sum.”

The present Article focuses on this latter, operational form of privatization. The state’s motivation in relying on the private sector in this way is a simple one—to find a way to house the growing inmate population while keeping costs down. And the prison-management companies themselves are equally financially motivated—the aim was, and is, to make a profit.

The first private entity formed to take advantage of this new business opportunity was Corrections Corporation of America (CCA), founded in Nashville in 1983. CCA’s founders had no experience in corrections, but from the start, the company’s management personnel were drawn from the public sector, including former state corrections commissioners, at least one former head of the federal Bureau of Prisons (BOP), and any number of former state prison wardens and superintendents. Wackenhut Corrections Corporation, the prison-management division of global security giant Wackenhut Security, Inc., entered the market soon after. Both CCA and Wackenhut were turning a profit by the late 1980s, and by the mid-1990s, they together controlled 75 percent of the American private prison market.

From the start, these companies faced a serious challenge, one that remains for any company trying to make money from running a prison. If the state is to reduce the cost of its prisons through contracting out to the private sector, the contract price must be less than the total cost the state would otherwise incur in operating the facility. And if private providers are likewise to make money on the venture, they must spend less to run the prisons than the contract price provides. For such arrangements to be remunerative for both parties, therefore, private prisons must be run at a considerably lower cost than the state would otherwise incur. At the same time, contractors must not allow either the quality of conditions of confinement or inmate safety to drop below existing levels; even staunch advocates of private prisons have insisted that “concern with cost savings should not outweigh considerations of quality.”
In practice, private prison providers have seemed little concerned with meeting this challenge. Instead, the anecdotal evidence suggests that contractors have prioritized economy above all else, with disturbing results for the inmates themselves. Consider, for example, CCA’s Youngstown, Ohio, facility. When the Youngstown facility opened in 1997, CCA filled the medium-security prison with prisoners from the overburdened Washington, D.C., prison system. The incoming D.C. inmates included a number of violent inmates classified as ‘maximum-security, high-risk,’ which CCA “reclassified” as medium security to fill the beds without having to equip the facility to handle maximum-security inmates. Over the next eighteen months, the Youngstown facility saw more than forty-four assaults and two fatal stabbings, including one inmate who was stabbed to death when a shortage of beds in the administrative segregation unit (a prison’s protective custody area) led prison officials to house the victim with two men who had been threatening his life.

At CCA Youngstown, economizing also took other forms. Former employees of the prison, for example, reported receiving a “rundown” by their employers, “saying two slices of bread per inmate costs this much. If you can cut corners here, it would mean a possible raise for us.” At Youngstown, even the “toilet paper was rationed,” to the point that “inmates were forced to go without it, using their bedsheets instead.”

Other incidents elsewhere suggest Youngstown is not unique for either its cost cutting or the troubling effects of such measures. This is not surprising, for efforts by private prison providers to cut costs even at the expense of inmates is the entirely predictable result of the existing structure of private prison contracts. Indeed, as I show, there is good reason to think that, where both the state and the contractor seek financial advantage, the challenge private prison contractors face—of running the prisons for less money than the state would otherwise pay without also bringing about a drop in the quality of prison conditions—cannot be met. There is, moreover, a further concern to which the use of private prisons gives rise, that fostering the private prison industry could create a powerful constituency with a financial interest in longer prison sentences, and the political clout to push sentencing policy in this direction, regardless of whether such punishments are consistent with the demands of legitimate punishment. These two possibilities lie at the root of the liberal critique of private prisons. But before this critique can be pursued, more must be said about the foundation of the legitimacy standard on which it rests.

II. A LIBERAL STANDARD OF LEGITIMATE PUNISHMENT

Legitimate punishment in liberal democracy has several components. Of these, two in particular bear most centrally on the legitimacy of penal policies and practices: the humanity principle, which obliges the state to avoid imposing punishments that are gratuitously inhumane, and the parsimony principle, which obliges the state to avoid imposing punishments of incarceration that are gratuitously long. In each case, gratuitous punishment is that which cannot be justified to all members of society under fair deliberative conditions. These principles reflect familiar liberal ideals: that society owes particular obligations of respect and consideration toward fellow human beings, especially those rendered helpless, dependent, and vulnerable by actions society itself has
undertaken, and that any violation of the liberty and dignity of citizens by the state demands compelling justification.

In what follows, I provide a foundation for these ideals in the theory of legitimate punishment in liberal democracy I have developed in greater detail elsewhere. Doing so grounds the intuitions informing the liberal principles of humanity and parsimony, and thereby enriches our understanding of the obligations incurred when the state punishes convicted offenders.

A. A Rawlsian Model of Legitimate Punishment

State punishment represents a dilemma for liberal democratic societies. For while punishment as a form of state power protects citizens from crime, it also represents the exercise of extremely oppressive force—at times even deadly force—by the state against its own citizens. A central challenge for any liberal theory is thus to establish the principles under which, in the name of criminal punishment, the state may legitimately burden, perhaps severely, the liberty, dignity, and bodily integrity of sovereign citizens.

How is the content of such principles to be determined? As I have elsewhere argued, following Rawls, if the exercise of the state’s power to punish in a liberal democracy is to be legitimate, it must be justifiable on terms that all those subject to this power would accept as just and fair under conditions of strict impartiality. Why conditions of strict impartiality? In a liberal democracy, all citizens are entitled to equal consideration and respect. All citizens, moreover, may be presumed to have an urgent interest in the greatest possible protection of what I have called their “security and integrity,” that is, security from assault on and interference with their physical and psychological integrity and well-being. These goods are fundamental to the exercise of individual freedom and self-development. They are also at great risk of violation by both crime and punishment. All citizens thus have an important stake in the terms on which state punishment is imposed on criminal offenders. But if these terms were established absent conditions of strict impartiality, there would be a danger that those parties with the most power and influence would simply choose principles of punishment that would most protect the security and integrity of people like themselves and do little to protect the security and integrity of society’s most vulnerable members. Indeed, the most powerful citizens might even choose principles of punishment that put the urgent interests of the most vulnerable citizens at risk, if doing so would benefit themselves in any way. Of particular relevance to the present project, for example, they might impose punishment regimes that burden the security and integrity of the most vulnerable in order that they themselves could benefit financially.

Applying Rawls's model of the “original position” with its “veil of ignorance” to the problem of punishment guards against this possibility. Behind the veil of ignorance, the parties selecting the principles of punishment know nothing of their own personal particulars or conception of the good. They can therefore only safeguard their own urgent interest in the greatest possible protection of their security and integrity if they choose principles that would also safeguard the like interest of all others. In this way, the strict impartiality of the modified Rawlsian model ensures that parties choosing principles of
punishment will consider the various options from all possible social positions—including that of society’s least powerful and most vulnerable members. This standard of strict impartiality thus ensures equal consideration and respect for the liberty, dignity, and bodily integrity of all sovereign citizens.

B. The Humanity and Parsimony Principles Derived

The question then becomes: what constraints on the state’s criminal justice policies would emerge from the deliberative model just described? Elsewhere, I identify several such constraining conditions. That analysis yields two principles, those of humanity and parsimony, which bear directly on the legitimacy of penal policies and practices. Space does not permit me here to provide full analytical support for this assertion, but brief consideration of the perspective of the deliberating parties in the original position should be sufficient to motivate the claim.

Behind the veil, the parties know nothing of their own social position or personal particulars, but they do know that they will have some conception of the good that they will want to realize. They also know that they are choosing principles of punishment for a partially compliant society, that is, a society with some measure of crime, where innocent people are sometimes wrongfully convicted and punished, and in which social goods are unjustly distributed. The parties will thus anticipate a threat to their security and integrity from both crime and punishment, and they will seek principles that best protect these goods. How are they to do so? Behind the veil, the parties would reason according to the “leximin” variant of the “maximin” rule. Maximin holds that under conditions of uncertainty, those wishing to maximize their prospects should assume that, once the veil is lifted, they will end up in the position of society’s worst off. Leximin then directs deliberators to select those principles that would guarantee the best possible result for the worst-off citizen who stands to be affected, and not to be “much concerned for what might be gained” by those who wind up in more fortunate positions.

The parties will thus select those principles of punishment that provide the greatest possible protection for the security and integrity of the worst off. This means the parties would not agree to principles that could compromise the security and integrity of the worst off in order that other better-off members of society might satisfy their less urgent interest in accruing financial advantage—an interest that is less urgent because it is unconnected to the protection of anyone’s security and integrity. This stance can be understood as constituting a priority rule—call it “the priority of the most urgent interests”—to govern the selection of the principles of punishment. Security and integrity are necessary preconditions for the exercise of all other basic liberties, prior even to material goods. The parties, consistent with the priority of the most urgent interests, would therefore reject any principles authorizing punishment that would only enhance anyone’s less urgent financial interests at the expense of the more urgent interest of the worst off in the protection of their security and integrity.

Both the humanity principle and the parsimony principle flow from this priority rule. Behind the veil, the parties cannot be confident that, once the veil is lifted and they enter society as citizens, they will not end up as either crime victims or targets of punishment.
They also know that incarceration represents a serious violation of the security and integrity of the target. They will thus choose principles of punishment that impose incarceration only when—and only to the extent that—doing so will maximize the security and integrity of the worst-off person who stands to be affected. To the extent that a prison sentence would worsen the condition of the target vis-à-vis that person’s security and integrity without improving anyone else’s condition vis-à-vis *their* security and integrity, it would be viewed as merely gratuitous and thus beyond the scope of punishments the state may legitimately authorize. Hence the parsimony principle, which prohibits gratuitously severe punishments.

To this point, humane punishment has been assumed. The parsimony principle is thus concerned exclusively with length of sentence. But what of inhumane punishment? Under some extremely limited circumstances, imposing inhumane punishment may be consistent with maximizing the security and integrity of the least-well-off person who stands to be affected. In the vast majority of cases, however, the imposition of any inhumane punishment would not satisfy these limited circumstances. And where it would not do so, it, too, would be merely gratuitous, and consequently illegitimate. Hence the humanity principle, which prohibits gratuitous inhumane punishment.

To identify the principles of legitimate punishment is no guarantee that the punishments actually imposed will in fact be legitimate. Many hurdles to effective implementation still exist. Perhaps chief among them is ensuring that the political process that translates the principles into policies remains unaffected by illegitimate influences. Such legislative-stage processes are as vulnerable as deliberation over the basic principles themselves to being skewed toward serving the interests, urgent or otherwise, of the politically powerful at the expense of the urgent interests of more vulnerable citizens. Ideally, to guard against any such abuses, parties deliberating at the legislative stage as to how to implement the principles of legitimate punishment would do so as if behind a “modified veil.” Such a veil would continue to screen out individuals’ knowledge of their personal particulars while allowing full access to the facts about society that are necessary to crafting meaningful policy.

In the real world, the legislative process falls somewhat short of this ideal. State officials, however, are still obliged to do what they can to secure the necessary conditions for legitimate punishment and to avoid taking steps likely to corrupt these conditions. This imperative may be thought of as an “integrity condition,” against which any criminal justice policy must be measured. Where legislators fail to satisfy this condition, the criminal justice system may come to lack integrity, a circumstance that could lead not only to illegitimate punishment but also to citizens’ widespread mistrust of the society’s criminal justice institutions. This danger, although certainly present where the issue is the humanity of conditions of confinement, is particularly salient in the parsimony context, where the issue is the severity of the sentences imposed. It is thus in the discussion of the parsimony principle that consideration of the integrity condition will most inform the analysis.
III. PRIVATE PRISONS AND THE HUMANITY PRINCIPLE

A. Understanding the Humanity Principle

The humanity principle, concerned with the conditions of confinement under which a given sentence is served, forbids gratuitous inhumane punishment. Inhumane punishments are those punishments imposed under conditions that degrade, humiliate, or otherwise seriously compromise essential aspects of the moral personhood of the target. I take it as uncontroversial that punishments of this sort would include those that subject targets to nontrivial deprivations of the basic necessities of human life—adequate food, clothing, shelter, medical care, and so on—as well as those that pose an ongoing threat of physical or sexual assault.

Inhumane punishment may not always be incompatible with the demands of liberal legitimacy in a partially compliant society. However, the circumstances under which such punishment might be legitimate are highly circumscribed, and at the very least would be subject to two main limiting conditions. First, where the state’s legitimate purposes can be achieved through either humane or inhumane forms of punishment, the state must impose only the former. And second, any inhumane punishment imposed must not be, in either duration or form, more severe than necessary to serve legitimate purposes. Each of these conditions is imposed for the same reason: any inhumane punishment beyond these points would be merely gratuitous and, therefore, illegitimate.

The limiting conditions on inhumane punishment required by the liberal perspective understand “gratuitous” punishment through the lens of the priority of the most urgent interests. Put more formally, no punishment that compromises the essential aspects of the target’s moral personhood may be imposed unless it can be reasonably certain and necessary to appreciably deter violations of the equally urgent interests of others who are as badly off as the incarcerated. It therefore follows that no inhumane punishment may be imposed in order to maximize the less urgent interests of anyone in their own financial gain, for any such inhumane punishment would necessarily be merely gratuitous.

How might this principle prohibiting gratuitous inhumane punishment be applied in the policy realm? For one thing, it would be an insufficient justification for inhumane treatment that money would be thereby freed up that could be put towards improving the prospects of free citizens. Instead, inhumane punishment that could otherwise have been prevented through greater financial investment would be justified on this principle only if it can be shown both that the money saved thereby is necessary to improve the condition of the worst-off person in society with respect to their most urgent interests, and also that this expected improvement is not mere speculation or vaguely anticipated future benefit, but is reasonably certain to result if the necessary resources are shifted away from the prisons. Where these twin conditions cannot be satisfied, the humanity principle obliges the state to spend the money necessary to prevent gratuitous inhumane punishment. The state is not required to spend more on inmates’ upkeep than is necessary to satisfy the minimum standard of the humanity principle, although in some cases prudence or other considerations may counsel doing so. Luxury accommodations are not required. Where, however, the state is faced with a choice between protecting prisoners from inhumane
conditions of confinement or funding some other appealing project, it is obliged to spend
the money to protect its inmates unless the competing project equally implicates the most
urgent interests of other citizens who are as badly off as the incarcerated.

B. Framing the Issues

In the private prisons debate, the dominant framework for assessing the desirability of
private prisons is what I have termed comparative efficiency. For this approach, the
motivating question is how private prisons compare with their public-sector counterparts,
and the central operating assumption is that if private contractors can run the prisons for
less money than the state without the quality of prison conditions falling below existing
levels, then the state should be willing to privatize.

It is important, however, not to let the impulse to comparative assessment distract from
the main purpose. My goal in this Part is not to vindicate one form of penal management
over another. It is instead to understand the structure and functioning of private prisons,
in order to assess the extent to which the state’s use of private prisons is consistent with
the demands of the humanity principle. Only with this understanding will it be possible to
see clearly the problems private prisons present, and to determine which of these
problems are unique to private prisons and which represent tendencies found in the prison
system as a whole.

This is not to suggest that comparison between public and private prisons is never in
order. Ultimately, the humanity principle is concerned with the health, safety, and well-
being of the incarcerated, and if these conditions turn out to be markedly better in one
prison form than another, this difference ought to matter to those committed to legitimate
punishment. But it is crucial that such comparisons not be premature, that they not
preempt a thoroughgoing analysis, and that they not be allowed to obscure the troubling
structural problems that plague both public and private prisons as currently constituted.
They must, moreover, be made on the basis of an appropriate measure: not cost, but
legitimacy, understood here in terms of the humanity of conditions of confinement.

Any adequate analysis of the structure and functioning of private prisons requires an
understanding of the motives of the two parties to the private prison contract. The private
contractors’ motives in both seeking and performing the contract are straightforward: to
profit from the venture. This very singularity of purpose is what is thought by many to
make private prisons so appealing, for private prisons seem to offer the state the
possibility of harnessing the profit motive to serve public ends, whatever those ends may
be. As for the state, in the American context, the central aim is to save money on the cost
of corrections.

The remainder of this Part considers the implications for the humanity of conditions of
confinement in a system in which both the prison contractor and the state are driven by
economic interests. Section C explores the likely effects of this incentive system on the
behavior of prison contractors, and Section D evaluates the efficacy of existing
accountability mechanisms. Having explored the incentive structure and the regulatory
landscape of private prisons, it will then be possible to make sense of one notable
difference between public and private prisons: the elevated rates of violence in private prisons. This phenomenon is discussed in Section E, which argues that, so long as the state’s use of private prisons is motivated by a desire to save money on the cost of corrections, private prisons are likely to be more violent and less humane even than state-run prisons.

This is not, however, to vindicate public prisons as currently constituted. Despite somewhat lower levels of violence, public prison conditions continue to be at odds in many respects with the demands of the humanity condition. Section F, therefore, looks to public prisons themselves, and finds that several of the most disturbing features of private prisons are also present in the public context. It then identifies some lessons to be learned from the analysis of private prisons for the penal system in general—lessons that, if heeded, could render conditions in all our prisons more consistently humane.

The first step in the analysis is to consider how private prisons might be expected to operate absent effective governmental regulation and oversight. The point here is not that no such regulation exists. But to move too hastily to asserting effective oversight short-circuits any possibility of fully understanding just what dangers are created by the state’s use of private prisons. Before it is possible to achieve this understanding, it is necessary to clarify exactly how the “profit seeking” behavior of private contractors might affect the running of the prisons were it allowed to operate unchecked.

C. A Thought Experiment: The Profit Motive Unconstrained

To determine how for-profit private prisons might be expected to operate absent effective restraints on contractors’ profit motive, consider the current structure of private prison agreements. Under the current system, the state agrees to pay a flat rate per inmate per day, and the contractor agrees to bear all the costs of running the prison. If the contractor is to make money, it must meet this contractual obligation for less than it earns from the state.

Private prison contracts thus contain a built-in incentive for the contractor to economize in two key respects. First, contractors will be tempted to reduce the amount spent on meeting inmates’ needs. In a prison, every aspect of inmates’ lives is dictated by the institution: when, what, and how much they eat; whether they get leisure time, adequate medical care, protection from harm, or access to rehabilitative or educational programming; the content and design of their beds and their cells; and when they shower and for how long. No detail of their lives remains unregulated. In a private prison, each of these aspects of inmates’ lives offers the potential for increasing profit margins. Absent effective checks, efforts on the part of private prison administrators to cut operational costs could thus lead to decisions that deprive inmates of basic human needs, a hallmark of inhumane punishment.

Second, profit-seeking contractors will be tempted to cut the cost of labor. As one industry observer explains, because two-thirds or more “of a prison’s budget goes to staffing and training,” private providers “must reduce expenditures in these areas if they are going to make a profit.” How might such cost cutting lead to inhumane conditions of
confinement? The more effective correctional officers are at maintaining a secure prison environment, the safer the inmates will be from the threat of physical assault. But guarding inmates requires constant interaction in a tense atmosphere with people who are bored, frustrated, resentful, and possibly dangerous. To protect inmates from harm and to ensure their own personal safety, correctional officers require training, experience, good judgment, and presence of mind. But when such officers are overworked and undertrained, or work in prisons that are understaffed, they are at a disadvantage in such a volatile environment and will thus be less effective at maintaining safe and secure prison conditions. Money-saving strategies that include “hiring fewer staff members, paying lower wages, and reducing staff training” thus increase the threat to inmates of physical assault, a further hallmark of inhumane punishment.

The foregoing account is likely to meet with two objections. First, some may object that it overlooks the potential of contractors to find innovative ways to reduce the cost of labor and other necessities so as to allow a comfortable profit margin without putting prisoners at risk. However, the existence of alternative avenues for profit making does not mean that contractors would not also seek to increase profit further in the ways predicted, were they able to do so without detection or penalty. There is, moreover, little evidence of cost-saving innovation in private-sector prisons. Nor, given the nature of prison administration, is there much scope for such innovation in this arena consistent with the humanity principle. Unlike Adam Smith’s butcher, brewer, or baker, private prison administrators are not dealing with inert materials. They are instead dealing in an extended and intimate way with human beings, whose treatment, if it is to be humane, requires constant attention and the careful exercise of discretion. Running a prison is thus necessarily labor intensive and affords little scope for more than marginal cost-saving innovation consistent with the humanity principle.

Second, some may object that the above account misunderstands the process of government contracting. Privately managed prisons, after all, exist only at the behest of the state. If the state wants to ensure a certain level of service provision, it need only specify its demands in the contract and hold the provider accountable. To achieve humane conditions of confinement, in other words, the contractual terms need only specify as much.

To some extent this is true. Where the standard of service to be provided can be specified in detail in advance, careful drafting can provide some protection from abuses. But with respect to many key features of prison life that are crucial from the humanity perspective—the use of force, health care provision, inmate classification, discipline, and inmate safety, among others—it can be difficult to specify in advance precisely how they are to be provided. To a significant extent, that is, private prison contracts are necessarily “incomplete,” meaning that the contractor’s obligations cannot be fully specified in the contract itself.

The inevitable incompleteness of private prison contracts raises two difficulties for efforts to rely on careful drafting alone to check contractor abuses. First, the necessarily vague character of incomplete contracts makes violations difficult to demonstrate and thus difficult to police. Second, because they are incomplete, prison contracts accord...
considerable discretion to contractors. This discretion comes in the form of what some economists call “residual control rights,” which carry “the authority to approve changes in procedure or innovations in uncontracted-for contingencies.” From the standpoint of prison administration, this allocation makes sense. Consider, for example the use of force. Plainly, it is not possible to spell out in advance every contingency within a prison that will require the use of force by correctional officers. Prison employees thus need discretion to use force when they think it warranted, for it is they who face an unpredictable environment and must make the hard judgments when potentially dangerous situations arise. Still, the extensive discretion necessarily lent by incomplete prison contracts to both line officers and prison administrators opens up space for these parties to use force against inmates in ways at odds with the demands of the humanity principle while still formally fulfilling the contract’s terms. Thus, even carefully drafted contracts cannot prevent many decisions by private contractors that might yield inhumane conditions of confinement.

There is, moreover, a further problem with relying on contract drafting alone to guard against possible contractor abuses. Even assuming the possibility of specifying contractual standards consistent with the humanity principle, the “hidden delivery” of prison services means that contractual violations may well go undetected. Imagine, for example, a contractual provision capping the number of assaults on inmates that may occur annually in the facility. To determine compliance with this provision, it is necessary for the state to have access to reliable data on such assaults. Yet private prison administrators are the ones who control access to this information, and they have a strong financial incentive to downplay the number of assaults that actually occur, particularly if this number exceeds contractual specifications. Thus, even assuming a contract that carefully delineated the maximum number of assaults, contractor control over the information necessary to effectively implement these contractual provisions could defeat this effort at regulation through contract.

D. Available Accountability Mechanisms and Their Limits

The claim so far is this: absent effective checks, the desire for profit will lead private prison contractors to cut costs in ways that will create or exacerbate gratuitously inhumane conditions of confinement. This claim is not a radical one. To the contrary, it reflects a basic assumption at the heart of the private prisons literature, one made by advocates and opponents alike, that without effective accountability mechanisms, privatization will lead to considerable reductions in the quality of prison conditions. The only difference here is that I have explicitly emphasized the costs of inadequate regulation in terms of the potentially inhumane treatment of inmates.

Properly channeled, the profit-seeking motive of private contractors may well allow states to achieve desired goals in terms of prison conditions without also creating the danger of contractor abuses. But to achieve such desirable results, effective regulation is indispensable. In what follows, I consider the four regulatory mechanisms most commonly introduced as evidence that effective checks on private prisons exist—the courts, accreditation, monitoring, and competition—and in each case explain why, under current and foreseeable circumstances, they are inadequate to the task.
1. The Courts

Arguably, any dangers private prison inmates face could be neutralized through lawsuits brought by them or on their behalf. Not only might abused inmates thereby get a remedy, but the threat of lawsuits and the accompanying possibility of major financial liability could provide incentives for private prison providers not to cut corners in ways likely to harm inmates. However, given the current state of the relevant law, the courts are not likely to provide a meaningful check on abuses in private prisons, notwithstanding the Supreme Court’s ruling denying private prison guards qualified immunity from Section 1983 actions.

Apart from a brief period in the late 1960s and early 1970s, judicial attitudes toward challenges to prison conditions have been marked by considerable deference to the judgment of prison officials. As a consequence, the constitutional rights of inmates have been interpreted extremely narrowly. For this reason, even instances of serious physical harm to inmates may not qualify for legal relief. Moreover, the mechanisms through which private prison providers might seek to save money could combine with the deferential standard of review under the Eighth Amendment to make it even less likely that private prison inmates could make out a successful constitutional claim.

Consider, for example, the use of force by prison officials against prisoners. For an inmate to have a viable Eighth Amendment claim against a prison official for use of excessive force, the inmate must show that the prison official acted “maliciously and sadistically,” with the intention to cause harm. So long as the prison official can make a showing that “the use of force could plausibly have been thought necessary,” the prisoner’s claim will fail. For example, even assuming that the corrections officers at a privately run jail in Brazoria, Texas, who “forc[ed] prisoners to crawl, kicking them and encouraging dogs to bite them,” engaged in this abusive treatment because they were insufficiently trained in less-abusive inmate control techniques, the prisoners themselves could have no constitutional recourse so long as the guards could plausibly claim to have thought their actions necessary to “preserve internal order and discipline.” Under these standards, private prison inmates suffering harm traceable to contractors’ inadequate investment in labor are even less likely to recover than public prison inmates: guards who are insufficiently trained may well resort to force more readily than guards with adequate training and experience, motivated in doing so not by a “malicious and sadistic” desire to cause harm, but by their own ignorance and fear.

Or consider the Eighth Amendment standard for prisoners alleging inadequate medical care. In *Estelle v. Gamble*, the Supreme Court held that for medical neglect of prisoners to rise to the level of an Eighth Amendment violation, prison officials must be shown to have acted with “deliberate indifference to serious medical needs.” To satisfy this standard, prisoners must show that prison officials actually knew of the health risk and failed to take reasonable steps to address the problem. It is not enough for the inmate to have told an official of pain or other physical distress; he or she must also show that the official actually “dre[w] the inference” from these facts of “an excessive risk to inmate health or safety.” Even under ordinary circumstances, it can be difficult for prisoners to make this showing. Add the profit motive to the picture, and the possibility of making out
a claim of Eighth Amendment medical neglect becomes even more difficult. Prison operators wishing to save money on medical care might, for example, create a deliberately unwieldy process for prisoners wishing medical attention, as has apparently been the strategy of Correctional Medical Services (CMS), a for-profit prison medical services company operating in prisons and jails in twenty-seven states. They might also hire medical staff of questionable competence, increasing the likelihood that conditions will go undiagnosed. Or they might institute treatment protocols of questionable efficacy that cost less than medically indicated methods. This last approach in particular might allow a defense that “reasonable” steps were taken even if they were ultimately ineffective.

Even assuming prisoners could demonstrate an Eighth Amendment violation, they must first get a hearing. Although no jurisdiction has ever warmly welcomed prisoner suits, the federal courts have traditionally been somewhat more receptive to prisoner claims than have state courts. However, the passage of the Prison Litigation Reform Act of 1995 (PLRA), intended by Congress “primarily to curtail claims brought by prisoners” under Section 1983, places severe limits on inmates’ access to the federal courts. In many cases, these burdens effectively prevent inmates’ constitutional claims from being heard in this forum at all. Not only does the PLRA explicitly limit the possible role federal courts might play in enforcing acceptable standards in penal facilities, but it also sends a strong message from Congress to the courts that they are to continue to give strong deference to prison administrators. These procedural hurdles, of course, also restrict court access for prisoners in publicly run facilities. But if the profit motive is a source of further potential abuse of prisoners in private facilities, these hurdles are that much more troubling when they prevent private prison inmates from gaining a hearing.

Private prison inmates do enjoy one doctrinal advantage over their counterparts in public prisons, an advantage that should, in theory at least, increase the likelihood that prisoners’ claims against private prison officials will succeed when like claims against public prison officials would fail. Under Richardson v. McKnight, private prison inmates filing Section 1983 actions need not overcome prison officials’ claims of qualified immunity. As a result, should private prison inmates be able to make a showing of unconstitutional treatment, private prison guards will be unable to escape liability on the grounds that the right they violated was not “clearly established” at the time of the violation.

Richardson, however, is unlikely to make much difference to private prison inmates. These inmates only have a true doctrinal advantage over inmates in public prisons when the right they are asserting has not previously been “clearly established.” If, however, prisoners are to succeed in vindicating constitutional rights not already clearly established, judges must add to the set of prisoners’ rights already recognized. And at present, there is little reason to expect federal judges to do so. Only during the late 1960s and 1970s did the Supreme Court seem willing to extend prisoners’ constitutional protections. And even during this period, the extent of this willingness was limited. The decades since, moreover, have seen a reinstatement of the “hands-off” attitude that predated that brief period of expansion. This recent retrenchment has been marked by a series of decisions paring back the rights articulated during the period of reform and creating new and substantial hurdles to the success of prisoners’ constitutional claims.
And these conditions are unlikely to change while public attitudes to incarcerated offenders remain as they are. Thus, the denial to private prison guards of the defense of qualified immunity is unlikely to benefit sufficient numbers of inmate plaintiffs to act as a meaningful check on the excesses of private contractors.

It might still be objected that, while courts are deferential to government officials, this deference is unlikely to extend to employees of for-profit prison-management companies. Private prison administrators and employees might thus not benefit from the culture of judicial deference to prison officials. This objection, however, misunderstands the role that judicial deference plays in prisoners’ rights cases. Recovery is difficult for prisoners, not because courts routinely show deference to the individual prison officials against whom suit is brought, but because, in the crafting of applicable constitutional standards, courts defer to the position and expertise of prison officials in general. Because the scope of prisoners’ rights under prevailing constitutional doctrine will be the same whether prisoners are housed in public or private facilities, private prison employees defending prisoner suits will enjoy the benefits of judicial deference to prison officials, whatever individual judges in specific cases may feel about the for-profit character of private prisons.

2. Accreditation

It is a standard requirement of state enabling statutes that private prison operators achieve and maintain official accreditation from the American Correctional Association (ACA), an independent “organization of correctional professionals dating to 1870.” The ACA sets standards governing every aspect of penal life and, on request, certifies the facilities that meet these standards to a satisfactory degree. The requirement that private prisons receive ACA accreditation is certainly desirable; indeed, in this regard, the private sector, having been forced to satisfy ACA standards, is ahead of many public-sector facilities, 20 percent of which did not have such accreditation in 2001.

Still, it would overestimate the effect of ACA accreditation to assume that this requirement sufficiently checks private-sector abuses. For one thing, ACA visits are highly structured, so that “certification [indicates] compliance with standards for only a brief period.” Moreover, the standards are largely procedural in character, generally satisfied by a showing as to “what the written procedures of the institution lay down as operational processes, rather than observing whether those processes in fact are followed.” Arguably, these problems could be resolved by an overhaul in the accreditation process, and such an overhaul would certainly be welcome. In its current form, however, the ACA is unlikely to undertake sufficient reform to ensure adequate protection against inmate abuses. For one thing, ACA officials are generally chosen from the ranks of experienced corrections officials. As a result, personal and professional relationships between ACA overseers and prison management are not uncommon, creating a common sympathy and sense of purpose that tells against both more meaningful standards and more rigorous enforcement. Moreover, the institutions being inspected “have” to pay for the whole procedure,” providing income on which the ACA is dependent for its survival. For this reason, “a degree of capture is likely.”
One could imagine a system of ACA accreditation that would serve as a meaningful check on declining prison conditions. Emphasis could be placed on ensuring conditions consistent with the humanity principle, prioritizing physical safety and the meeting of basic human needs. To be successful, however, any such reform would need the backing of ACA membership and state officials alike. Moreover, more frequent and effective monitoring would be required, which is both expensive and itself susceptible to the problems of capture. These problems are not insurmountable ones. However, where the state’s aim in privatization is to save money, little progress may be expected toward effective ACA standards that satisfy the humanity principle.

3. Monitoring

As John Donahue has observed, “full, effective monitoring [of private prisons] is a tall order.” Why is this so? In the prison context, the hidden delivery of the contracted-for services means that the contract is fulfilled away from the scrutiny of the buyers—in this case, the state. Prisons are often large, sprawling institutions, housing anywhere from several hundred to several thousand inmates. At any time in a given facility, therefore, scores and perhaps hundreds of employees are operating in a volatile environment, shielded from public view.

The call for monitoring is the usual response when concern is expressed regarding the possibility of abuses by private prison contractors. Yet, available data reveal good reason to doubt the efficacy of the monitoring regimes in place to oversee contractual compliance. The most comprehensive survey on the question was conducted in December 1997. This survey found that, of the twenty-eight state and federal government agencies then in the midst of “active contracts with privately operated [penal] facilities . . . twenty reported using monitors in addition to contract administrators,” suggesting that fully eight agencies used no monitoring at all. The twenty agencies that reported using on-site monitoring provided survey information for ninety-one separate contracts. Of these, forty-six—slightly over half—reported having monitors on-site on a daily basis. The remainder had monitors on-site weekly (five), monthly (sixteen), quarterly (ten), “on an ‘as needed’ basis or on an annual or semi-annual basis” (nine), with three contracts conducting all their monitoring off-site.

What should be made of this data? Given the enormity of the task of overseeing contractual performance under circumstances of “hidden delivery” in crowded and bustling institutions, it seems plain that systems under which monitors make only occasional on-site visits are inadequate to the task—even assuming, as the data suggest, multiple monitors per visit. As the authors of the study themselves note, “[w]here monitoring is so limited, it is unlikely that contracting agencies are able to provide more than a cursory assessment of the contractor’s performance.” Certainly, those contracts that provide for full-time on-site monitors are an improvement over those that allow for only occasional visits: the average permanent on-site monitor spends an average of 7.25 hours per day, working five days a week, in the monitored facility. But still, given the scope of prison contracts and the range and extent of the interactions and activities within any given prison, it seems unlikely that comprehensive and meaningful oversight can be achieved by a single monitor spending an average of thirty-six hours a week on-site.
It is theoretically possible that a comprehensive system of contractual oversight could check the temptation of contractors to cut costs in ways likely to harm inmates, if the contractors actually believed that the decisions made by their employees would be observed and recorded by monitors committed to enforcing the terms of the contract. But this possibility provides little comfort if—as the data suggest—no such comprehensive system actually exists.

Why are existing monitoring systems so inadequate? The answer is at least in part financial. Monitoring is necessarily labor intensive and therefore expensive, requiring an investment that states—which turned to privatization to save money—are not eager to make. States may try to pass the cost of monitoring onto the contractor, but such efforts are ill-advised. Unless the contract specifies the amount contractors must spend in this regard, contractors’ interest in cutting costs—not to mention their interest in reducing the effectiveness of monitors in exposing contractual violations—will likely lead to an investment too small to serve the purpose. And were the contract to stipulate the expenditure of an amount sufficient to ensure effective monitoring, it could well erase the possibility of any profit margin for the contractor. Private prison providers already operate on extremely narrow profit margins, so if the state is to have any contracting partner at all, imposing such stipulations would also necessarily drive up the contract price for the state. There is no way around it: if monitoring is to be effective, the state must bear the cost.

Even assuming adequate financial investment on the part of the state, however, there remains a further obstacle to the effective monitoring of private prisons: the risk of “agency capture.” Agency capture occurs when “regulators come to be more concerned to serve the interests of the industry with which they are in regular contact than the more remote and abstract public interest.” The worry here is that monitors will become too closely aligned with the facility being monitored, leading them to overlook or miss altogether evidence of abuse.

Although relations between state-employed monitors and private prison management are likely to be closest when monitoring is carried out by a full-time on-site inspector, opportunities will exist for a rapport to develop between inspectors and contractors, whether the monitor is permanent or makes only periodic inspections. These actors, generally drawn from the same pool of corrections professionals, all share a common interest, knowledge base, professional community, and perhaps most importantly, a sense of the difficult challenges involved in running a prison and a concomitant sympathy with the perspective of prison administrators. Such a rapport can orient the monitor toward the interests of the contractor, making it less likely that contractual performance will be challenged. Equally significantly, the “revolving door” between state agencies and private providers can “create [a] subtle conflict of interest,” as monitors who might at some point seek to move from public to private employment try to avoid alienating potential future employers in the course of performing their current responsibilities.

Effective monitoring thus appears to have two key requirements: sufficient financial investment, and a commitment to overcoming the risk of agency capture. Even still, the scope of activity within the prison and the hidden delivery of prison services may limit
the likely effectiveness of any monitoring scheme. Notice, moreover, that there is a
tension between these requirements, in that the more time monitors spend on-site, the
greater the risk of agency capture. Although this tension need not undermine the
possibility of effective monitoring, it does further indicate the limits of monitoring as a
possible check on contractor excesses.

4. Competition and the Threat of Replacement

Even assuming a contract could be drafted with sufficient specificity to reflect the desired
results and even if an effective system of monitoring were in place, prison contractors
need not fear exposure of noncompliance absent a credible threat of replacement. Plainly,
the states need to house their prisoners somewhere. Contractors know this, and know too
that states face great obstacles to finding suitable alternative accommodations for their
prisoners. They will therefore understand that, notwithstanding threats to this effect,
contractual noncompliance need not necessarily mean a loss of the contract.

As Donahue points out, “[p]erfect competition—many alternative suppliers, ease of entry
and exit, full information, and so on—is out of the question here.” But is the field of
competition good enough to ensure a meaningful threat of replacement in the event of
nonperformance? At least three characteristics of the private prison “market” raise
questions as to the likely efficacy of such a threat in ensuring ongoing quality of service.
First, as Justice Scalia points out in his dissent in Richardson v. McKnight, the only
buyers in this market are public officials, spending “other people’s money.” Consequently,
factors other than quality of service are liable to influence the judgment of
whether to cancel or renew a contract—for example, politicians’ need to secure future
campaign contributions or the business or personal connections between politicians,
corrections officials, and firm management. Second, there is a relatively small number of
viable industry participants with the experience, resources, and infrastructure necessary to
make a bid. For this reason, although a number of smaller companies have sought to
break into the market, the industry continues to be dominated by a very few major
players. The limited pool of competitors can undermine the force of threatened
replacement even in the event of inadequate performance on the part of the contractor.
Third, the dependence of the government on the initial provider is compounded by the
obstacles to the state’s resuming the operation of a privatized facility: it would face high
start-up costs, especially if the current facility were owned by the contractor, and possible
litigation arising out of the termination of the contract. In rescinding a private prison
contract, the state is therefore likely to wind up spending more on corrections than it had
before privatizing. It may thus “be cheaper for the state to accept some contractor abuses
than to remedy them by resuming state operation.”

Available evidence confirms that, absent both political pressure to replace abusive or
otherwise ill-performing contractors and a willingness to bear the financial cost of such
replacement, the state is unlikely to act on the threat of rescission. It is not that state
agencies never replace contractors in the event of noncompliance; states experimenting
with privatization have rescinded a number of private prison contracts after contractor
abuses came to light. But what seems to be required for such cancellation are conditions
sufficiently objectionable to trigger a public outcry, an effect that generally occurs under
limited circumstances: either the inmates experiencing abusive conditions are housed by private prisons out of state, or the exposed conditions are extremely egregious.

As to the former, at least six states have cancelled contracts “involving the shipment of [their own] prisoners to private prisons in another state” following allegations of “vendor violations.” Among them was Missouri, which pulled eight hundred of its prisoners from three Texas jails managed by Capital Correctional Resources, Inc. (CCR), after a leaked videotape showed Missouri inmates in CCR’s Brazoria facility “forced to crawl on the floor, shocked with electric prods, [and] bitten by police dogs.”

As to the latter, in 1995, for example, the INS cancelled its contract with Esmor Correctional Services for the operation of its Elizabeth, New Jersey, facility after it came to light that Esmor’s practices included the “cutting [of] financial corners” on food, so that at some Esmor facilities “there were often only 30 meals to feed 100 inmates, because Esmor did not want to pay for more.” Detainees were also denied such essentials as clean underwear and sanitary napkins, and Esmor even charged them “for lost eating utensils, clothing and drinking cups.” Investigations into the New Jersey facility also revealed that guards who were ill-trained, overworked, and outnumbered had routinely abused inmates physically, “shackled them during visits [and] placed them in punishment cells for little documented reason . . . [as] part of a systematic methodology designed by some Esmor guards as a means to control the general detainee population.”

To take one further example, in 2004, in the wake of extensive reports of abuse, the state of Louisiana closed its Tallulah Correctional Center for Youth. Tallulah had been operated by Trans-American Corporation, a company run by a local businessman whose father was “an influential state senator.” At the Tallulah facility, inmates had “regularly appear[ed] at the infirmary with black eyes, broken noses or jaws or perforated eardrums from beating by the poorly paid, poorly trained guards or from fights with other boys.” One inmate suffered regular beatings from guards, and after fifteen months, a judge ordered that he be released so he could receive medical attention. By this time “his eardrum had been perforated in a beating by a guard, he had large scars on his arms, legs and face, and his nose had been broken so badly that he [spoke] in a wheeze.” Trans-American “scrimped on money for education and mental health treatment . . . to earn a profit,” and meals at the facility were “so meager that many boys los[t] weight [and c]lothing [was] so scarce that boys [fought] over shirts and shoes.”

The willingness of state agencies to cancel contracts under these limited circumstances suggests that the threat of replacement may serve to check at least some contractor excesses, particularly when the private prison is located in another state. But taken as a whole, the evidence suggests not a readiness to rescind contracts when there is evidence of widespread abuse but reluctance on the part of states to do so even in the face of long-term concerns with prison conditions. Wisconsin, for example, waited five years after allegations first surfaced of physical and sexual abuse of Wisconsin prisoners in a CCA-run prison in Tennessee before announcing its intention to cancel its contract and bring its prisoners home, although the allegations had been confirmed by a team of Wisconsin state investigators shortly after being raised. And Louisiana’s Tallulah Correctional
Center for Youth, which the state finally closed in 2004, had by that time seen allegations of severe abuse of prisoners for a full decade.

To make the threat of replacement meaningful, legislators must commit to bearing the cost when the evidence of abuse suggests the need to do so. Yet, where the state’s priority is saving money, this willingness is unlikely to be forthcoming absent extreme circumstances and a public outcry. As a consequence, the threat of replacement cannot be expected to deter any but the most extreme abuses.

E. Private Prisons: Problems and Prospects

The foregoing survey suggests that, although existing oversight and accountability mechanisms are not wholly ineffectual, they fall far short of providing adequate safeguards against prisoner abuse. Ideally, private prisons would allow states to harness “[the] willingness and ability [of the private sector] to innovate in pursuit of profits” within a regulatory structure that effectively checked any efforts by contractors to save money in ways likely to put inmates at risk. But this is not the regime currently in place.

Instead, absent the restraining power of effective regulatory and oversight mechanisms, private prison contractors have acted largely as earlier predicted. That is, they have sought to increase their margins by considerably reducing their labor costs, systematically cutting salaries and benefits to employees, and underinvesting in training. They have done so, moreover, without fear of either contravening statutory civil service protections or meeting collective resistance from their workers. Not only are the employees of private prisons not state employees, a fact that allows their employers to set contract terms with minimal restrictions, but private prison employees are also not generally union members. As employees of private companies, guards in private facilities are not eligible for membership in the American Federation of State, County, and Municipal Employees (AFSCME), which has so effectively represented publicly employed correctional officers nationwide. Nor have private prison guards tended to form their own unions, as “private correctional companies make every effort not to employ unionized workers and not to let their workforce join any union.”

Nor have predicted innovations in prison management through privatization come to pass. Instead, the private prisons of today function very much like public prisons, only with a cheaper labor force. Private prisons thus generally exhibit all the particularized characteristics that make public prisons dangerous places: the considerable discretion and power conferred on guards; the fear on all sides; the simultaneous monotony and high pressure of the prison environment; inmates’ possible proclivity to violence; and the relative social and economic disempowerment of prison guards, who do a difficult job in a tense and dangerous environment and for whom power over prisoners constitutes both a rare perquisite and an outlet for frustration. But in addition, private prison employees are likely to be less qualified (because less well remunerated) and less well trained than their public-sector counterparts.

Given this situation, it seems likely that private prisons as currently constituted would turn out to be more violent places than their state-run counterparts. And in fact, although
much of the available data is inconclusive regarding the overall quality of conditions in private prisons as compared with public facilities, meaningful data do exist showing elevated levels of physical violence in private prisons.

For example, in 1997, researchers at the U.S. Department of Justice Bureau of Justice Assistance (BJA) surveyed private prison operators and received responses pertaining to sixty-five of the eighty private correctional facilities then in operation in the country. They then compared this information to comprehensive data on public prisons nationwide. Comparing the number of “major incidents,” including “assaults, riots, fires and other disturbances” in the public prisons over twelve months with those occurring in private facilities over the same period, the survey found a greater number of such incidents per one thousand inmates in the private prisons: 45.3 per 1,000 inmates in public prisons, as compared with 50.5 in private facilities. When inmate assaults were taken alone, the disparity was even more marked: 25.4 per 1,000 inmates in publicly run facilities, as compared with 35.1 in private prisons.

These data, moreover, do not account for the greater proportion of maximum-security inmates in publicly run facilities—19.8 percent as compared with only 4.6 percent in private facilities. Maximum-security prisoners are so classified because they are considered a much greater security risk and are thus likely to be more violent than prisoners with lower security classifications. One should thus expect public prisons, which house a higher proportion of maximum-security inmates, to be more violent than private ones. That private prisons are more violent than public ones despite the lower security classification of private prison inmates suggests that particular violence-fostering forces are at work in private prisons that are not present in public prisons, or at least not present to the same degree. This hypothesis is reinforced by the picture that emerges once the data are adjusted to compare only “the medium and minimum security public facilities with the same type of private facilities.” Here, the difference is even more pronounced, with 29.6 “major incidents” per 1,000 inmates at public prisons, as compared with 48.0 in the private facilities. For inmate assaults taken alone, with the adjustment for security classification, public prisons had 20.2 assaults per 1,000 inmates, as compared with 33.5 in private facilities.

The data on staff assaults likewise changed notably once the classification levels were taken into account. Private prisons did slightly better than public prisons on this measure when the full complement of publicly held maximum-security inmates was included: per 1,000 inmates, there were 12.7 assaults on staff in private prisons as compared with 13.8 in public prisons during the twelve-month period studied. However, when the data were recalculated to include only medium- and minimum-security inmates, researchers found 8.2 such assaults per 1,000 inmates in public facilities, as compared with 12.2 in private prisons.

Other studies have also found elevated violence levels in private prisons as compared with public ones. For example, according to Judith Greene, a New York-based expert on private prisons, a comparative study of “serious incidents” in public and private facilities in Oklahoma over a three-year period found that “private prisons recorded more than twice as many incidents as public ones.” Similar findings were also made in an earlier
study commissioned by the Tennessee Department of Corrections (TDOC). The TDOC study compared two public prisons and one prison run by CCA. Although the study authors claimed to have found no significant differences among the prisons in terms of quality, the empirical data on which this conclusion was based tell a different story. In particular, over the fifteen months studied, “the private prison reported significantly more (214) injuries to prisoners and staff, compared to 21 and 51 for the two state prisons respectively,” and “[t]he private prisons also reported 30 incidents of the use of force [against inmates by guards], compared with 4 and 6 respectively for the state prisons.” As with the BJA study above, inmate characteristics were not consistent across the three Tennessee facilities, with each facility housing “quite different types of inmates in terms of the socio-demographic characteristics reported, age and race, criminal history and custody classification.” That this is so, however, only strengthens the point, for prisoners assigned to the care of private prison providers tend to be the “cream of the crop,” those thought to be less inclined to violence or other forms of troublemaking. One should thus have expected fewer incidents of injuries and use of force in Tennessee’s private prison, rather than the other way around.

Certainly, nothing in the foregoing discussion goes to show that the state’s use of private prisons could never satisfy the humanity principle. What it does show is that, when the state looks to privatization to save money on the cost of corrections, there is reason to expect conditions of confinement to fall below even that level of quality and safety that can be reasonably expected of those charged with the difficult task of running the prisons. When the state’s aim is saving money, it will be unwilling to undertake measures that will substantially raise the cost of privatization, even when doing so could arguably ensure more meaningful protections for vulnerable inmates. So the state will invest minimally in monitoring contractual compliance, placing perhaps one full-time monitor at each site, despite the arguable need for a full-time team of monitors if the effort is to be at all effective. When money is the state’s primary concern, it will hesitate to rescind contracts even when evidence of abuse is considerable, fearing the costs such a move would entail. It will also forbear from specifying contractual terms requiring private contractors to provide minimum levels of staffing and training for private prison guards and stipulating the salaries and benefits to be paid to them. Doing so would only increase the cost of contracting to the state and would, moreover, greatly tie the hands of contractors, for whom cutting labor costs is the central available means to keep expenses down.

It is possible that the hazards private prisons pose under these circumstances might be mitigated considerably were society committed to satisfying the demands of the humanity principle and willing to pay the cost of effective regulatory tools. Still, even assuming such a commitment, obstacles would remain to eliminating gratuitously inhumane treatment in private prisons. Contracts would still be incomplete and would continue to accord residual control rights to private prison administrators and guards, thus allowing scope for abuses. Even assuming a public commitment to adequate oversight, many aspects of prison life would inevitably still go unobserved. And though one might contemplate a change in the public sentiment regarding the cost of corrections, contractors themselves will still be motivated by the desire for profit. This means that even under the altered circumstances contemplated, the contractors’ interests would
remain at odds with those of the humanity principle, continuing to place a burden of particularly rigorous oversight on the state.

F. Public Prisons: A Satisfactory Alternative?

To judge private prisons from the perspective of the humanity principle, it is not enough to consider merely the idea of private prisons. Instead, it is necessary to examine prison contracting as a practice, within the regulatory context in which private prisons actually operate. Conducting this contextual analysis has made plain that, at least as currently structured, private prisons pose an appreciable danger to the possibility of legitimate punishment.

Yet public prisons, too, will invariably fall short when measured against the ideal the humanity principle represents. Given conditions in publicly run prisons and jails, it would be absurd to suggest otherwise. The question then becomes, of what relevance is the sorry state of many public prisons to the present discussion?

Were comparative efficiency the operative framework here, drawing attention to existing conditions in public prisons would serve as a rejoinder: yes, conditions in private prisons are at odds with the demands of the humanity principle, but so are conditions in public prisons—the implication being that, given the shortcomings of public prisons, private prisons may still be preferable. But liberal legitimacy rejects the “either/or” approach of comparative efficiency. It aims not to champion the least bad alternative, but instead to understand how and why existing prisons and jails, public and private, fall so far short of the ideal. For liberal legitimacy, therefore, the current state of public prisons represents not a rejoinder to the foregoing critique of private prisons, but rather an occasion for asking whether the insights gleaned from that critique help also to explain failings in the public system.

That there is much to learn about public prisons from a study of private prisons only makes sense. Although the privatization of corrections is generally treated in the private prisons literature as a radical departure from prevailing penal practices, prison privatization represents neither an isolated nor an aberrant approach to punishment. It is instead the logical extension of practices that are standard fare in the prison system as a whole. Of these practices, two in particular bear emphasizing: (1) the widespread use of private contractors to provide key prison services at a cheaper cost than the state would otherwise pay, and (2) the delegation to correctional officers of considerable discretion—and thus, considerable power—over vulnerable and dependent prisoners absent mechanisms adequate to check possible abuses.

The foregoing discussion suggests that the risk that private prisons will be unsafe and inhumane stems directly from these two practices. Yet both of these practices are also standard components of state-run prisons across the country. As to the first, virtually every corrections facility in the country contracts out to for-profit providers for at least some necessary services, including everything from food services to medical, dental, and psychiatric treatment to rehabilitative and educational programming, garbage collection, and even inmate classification. The state’s primary motivation for such contracting is its
potential to cut the cost of corrections to the state; for the majority of contractors, as in the case of private prison providers, the aim is to make as large a profit as possible. And although perhaps some services—say, garbage collection—can be carried out without having an impact on prison conditions, most others directly affect the well-being of prisoners.

And as to the second, in terms of correctional officers’ discretion in dealing with prisoners, there is little if anything to distinguish public from private. Prison officials in public and private prisons alike have direct control over all aspects of prisoners’ day-to-day lives, the circumstances of which are well hidden from public view. Furthermore, the mechanisms in place to check potential abuse in the public prisons are either identical to those that apply in the private context, or else, despite differences, are just as likely to be inadequate.

With respect to the courts, aside from the narrow Richardson exception, the legal standards previously canvassed, both substantive and procedural, apply equally to public and private prisons. The same holds for the certification standards of the ACA, which does not distinguish between public and private prisons when assessing facilities for accreditation purposes, and thus judges each on the same basis. As for the threat of service-provider replacement in cases of poor performance and ongoing monitoring of internal prison affairs, although the structuring regimes differ in each case, these mechanisms appear no more effective at constraining abuses in the public sphere than in the private.

With regard to competition and the threat of replacement, because the state has a monopoly over prison administration, there is no possibility that a dysfunctional public prison will be taken over or replaced by an alternate provider. And as to monitoring, although many states provide by statute or administrative regulation for regular inspection and oversight of corrections facilities, the stipulated requirements tend to be minimal. In California, for example, the Office of the Inspector General (CA OIG) is required by statute to audit each prison within a year of a new warden’s appointment, and all facilities once every four years. Although available reports suggest that the CA OIG does a thorough job and provides detailed and useful findings, the infrequency of the audits suggests that this oversight scheme can have only limited effect. At the same time, even where monitoring requirements seem on paper to be relatively stringent, they do not necessarily translate in practice into effective means for identifying and checking possible abuses. In Tennessee, for example, the Commissioner or Deputy Commissioner of Corrections is required to visit each state prison once a month “to determine whether the laws, rules and regulations are duly observed . . . and the convicts properly governed.” Yet my efforts to turn up records confirming these monthly inspections were unavailing, and the commissioner’s own annual review suggests that inspections are conducted not monthly but annually. And even these annual inspections seem unlikely to affect the quality of conditions of confinement, given the emphasis the commissioner places on the “cost efficient” character of this monitoring scheme; as has already been seen, an emphasis on cost savings in the context of monitoring tends to detract from, rather than enhance, the success of the effort.
Thus, public prisons, too, contract with for-profit providers for the provision of essential prison services as a cost-saving measure. And in public prisons, too, correctional officers enjoy considerable power over prisoners absent effective oversight mechanisms. It should thus be unsurprising that, in terms of day-to-day structure and functioning, private prisons operate pretty much like public prisons—and that the conditions in each are far from safe or humane.

What lessons, if any, does the foregoing analysis of private prisons offer for those wishing to increase the humanity of prison conditions generally? For one thing, it suggests that prison officials should be wary of contracting out prison services, even on a smaller scale, to any entities that promise to reduce the cost of providing essential prison services in exchange for the chance to make a profit for themselves. This caution extends not just to contracts for running whole prisons, but also to the contracting out of discrete prison services like food service, medical, dental and psychiatric care, rehabilitative programming, and inmate classification. As has just been seen, absent effective checks, one can expect for-profit contractors to cut costs even at the expense of inmates. Creating disincentives to this behavior is therefore crucial. But ensuring meaningful oversight and accountability costs money, and any time the states contract out to reduce their prison budgets, state officials are going to be reluctant to spend what it takes to ensure prisoners’ ongoing security and well-being. This set of dynamics means that contracting out even discrete prison services to for-profit contractors when the state’s goal is cost cutting is a recipe for seriously compromised conditions of confinement.

Experiences with prison health-management companies bear out this caution. To take just one example, in 2003 alone, CMS took in over $500 million contracting with prisons in thirty states to provide medical care for inmates. Although the company is extremely secretive, investigations into company practices have revealed a litany of stories of inmates who died or suffered serious long-term disability because of treatment delayed or denied, of staff—doctors and nurses—being hired despite their having been suspended from the practice of medicine or otherwise disciplined by the medical boards issuing their licenses, and of policies deliberately designed to minimize the amount of medical care ultimately provided to prisoners in need of treatment.

And contracting out is not the only practice that bears scrutiny for the reasons just explored. It is also necessary to scrutinize carefully any independent efforts on the part of state officials to make publicly run prisons financially self-sustaining or to run them at a profit. Meeting inmates’ needs and ensuring their safety is expensive and requires a certain minimum investment. Yet state corrections officials, too, face pressures to reduce the cost of running the prisons. Depending on the approach adopted to achieve this end, these efforts could well pose a risk of serious harm to inmates.

The foregoing analysis of private prisons also indicates that dangers may exist whenever individuals with their own motives and interests are given wide discretion over the lives of inmates. Given the power wielded by all correctional officers, public or private, all prisons could benefit from more intense efforts to monitor meaningfully and to enforce standards of confinement consistent with the demands of the humanity principle. All prisons, for example, ought to be required to secure ACA accreditation, although ideally
these standards would be made more rigorous than they currently are. There ought to be benchmark standards of quality and humanity that apply to all prisons, set by state departments of corrections themselves—or if this arrangement would create too great a conflict of interest, then by an independent body. This body could also be charged with monitoring compliance in a regular and ongoing way. In the public sector, too, such an arrangement would carry a risk of agency capture, but absent the lure of a high-paying job with the entity being monitored, monitors overseeing the performance of state-run prisons might at least be somewhat less vulnerable in this regard. And efforts should be made to render the courts a more meaningful channel for ensuring the accountability of prison administrators. Such efforts might include clarifying (and even strengthening) via statutory enactments the particular duty of care that state officials owe inmates, providing inmates seeking to enforce these standards with more meaningful rights of access to the courts, and establishing liability rules more likely than prevailing Eighth Amendment standards to yield humane conditions of confinement. Such changes to prevailing legal standards would benefit inmates in any prison, public as well as private.

As these proposals indicate, viewed from the liberal perspective, there is a great benefit in shifting the focus of the private prison debate from efficiency to the humanity of conditions of confinement. Doing so allows us to transcend the inadequate baseline of current prison conditions and to consider how the system as a whole, public prisons as well as private, might better measure up against society’s obligations to those it incarcerates.

These proposals do leave unaddressed one accountability mechanism favored by private prison advocates, that of competition and the threat of replacement. The absence of these forces in the public context, a product of the state’s monopolistic status as the sole prison provider, is one reason some advocates of private prisons offer for their preference for privatization. On this view, the fear of losing one’s contract to a competitor is supposed to exert a disciplining force not available in the public context. Yet the foregoing discussion suggests that this threat is far less effective at curbing abuses than is often believed. It also suggests the danger of relying on this threat to ensure adequate conditions of confinement; even if a given prison contractor is replaced, there is no guarantee that the replacement will do the job any better. Indeed, there is reason to think they will not do so, given that the same regulatory structure that failed to constrain the previous provider will still govern. Moreover, having replaced a provider once, the state will not be eager to do so again. The lesson here applies equally to public prisons as to private ones. If the penal system is failing, changing the logo on the letterhead or the nameplates on the doors will not solve existing problems. When the system is failing, the system itself requires serious reconsideration.

IV. PRIVATE PRISONS AND THE PARSIMONY PRINCIPLE

A. Understanding the Parsimony Principle

As conceived of here, the parsimony principle speaks not to conditions of confinement, but to the length of prison sentences. This principle obliges the state to avoid imposing sentences that are gratuitously long. I have argued elsewhere that this constraint requires
that punishments be no more severe than necessary to appreciably deter offenses causing harm of equal or greater severity. But one need not accept this particular interpretation of the parsimony principle to agree with the more general point that to be legitimate, punishment can be of no greater severity than can be justified to all under fair deliberative conditions.

This more general version of the principle is all that is required to motivate the analysis of private prisons from the parsimony perspective. Whatever one’s view of the justification for punishment that would be accepted under fair deliberative conditions, any view satisfying this requirement would recognize the burden that punishment places on the urgent interests of the targets and would thus authorize the state to punish convicted offenders only when interests of equal or greater urgency would thereby be served. I therefore take it as uncontroversial that, whatever one’s particular view as to the justification for punishment that would be accepted under fair deliberative conditions, all would reject the idea that punishments of incarceration might be imposed on some members of society in order that others might benefit financially.

From the parsimony perspective, any such punishments would be considered gratuitous, and therefore illegitimate. Yet the privatization of prisons seems to raise this very concern—that the delegation of responsibility for prison administration to private, for-profit contractors could result in punishments imposed so that some members of society might increase their net worth. To this extent, the state’s use of private prisons would violate the parsimony principle.

Efforts to ensure that prison sentences satisfy the parsimony principle face obstacles not present in the humanity context. With very few exceptions, violations of the humanity principle can be verified against existing prison conditions. People may disagree as to whether particular conditions are inhumane, but the ultimate judgment will turn on the nature of the conditions themselves. To identify violations of the parsimony principle, in contrast, it is not enough to look to the sentences themselves, for there are no qualitative differences between legitimate punishments and gratuitous ones. All are measured by incarceration for a term of years, and what for one offender will constitute gratuitous punishment will be legitimate when applied to another. In the parsimony context, it is the reason why a particular sentence was authorized that matters. The determination as to whether a particular punishment is illegitimate is thus necessarily far more speculative in the parsimony context than in the humanity context. Moreover, it is the process of fixing sentences rather than the particular sentences themselves that is the primary focus of scrutiny.

The parsimony principle requires at the very least that punishments not be imposed on some members of society in order that others might benefit financially. It is thus crucial that the process of shaping sentencing policy and fixing particular sentences be untainted by illegitimate influences. In practice, of course, this process will inevitably fall short of this ideal. As a consequence, even legislators committed to realizing the parsimony principle cannot be confident that either the sentencing policies they authorize or the punishments imposed pursuant to those policies will satisfy this standard. Such legislators thus have an added burden. Not only are they obliged to do all they can to ensure the
conditions necessary if criminal punishment is to be legitimate, but they must also avoid taking steps likely to corrupt these conditions. This “integrity condition” is designed to ensure that the process of crafting sentencing policies is as isolated as possible from any illegitimate influences. When this condition remains unsatisfied, the criminal justice system itself may come to lack integrity, a circumstance that could lead to the imposition of illegitimate punishments. It could also, moreover, lead to citizens’ mistrust of the judgments and actions of government institutions. Citizens may be expected to respect the authority of the state only when they trust that state power is exercised for legitimate reasons justifiable to all. When the process of setting criminal penalties is perceived to lack integrity, citizens would reasonably lack confidence that punishments imposed on convicted offenders were consistent with such reasons. Satisfying the integrity condition in the parsimony context is thus important for two reasons: it increases the likelihood of legitimate punishment, and it secures public trust in the process.

In what follows, I consider the possible parsimony concerns raised by the state’s use of private prisons. In contrast to Part III, the present discussion introduces considerations that have not previously been addressed in any systematic way in the private prisons literature. It is thus necessarily more speculative and suggestive than Part III. As with the discussion in Part III, consideration of the state’s use of private prisons from the perspective of the parsimony principle suggests a critique, not only of the state’s use of private prisons, but of the penal system as a whole. In this case, however, it is not merely particular practices, but the whole process of policymaking in the criminal justice context that the analysis of private prisons calls into question.

**B. Influencing Time Served from the Inside: “I’m the Supreme Court”**

The length of the sentence a convicted offender officially receives is established by the sentencing judge within the terms set by the legislature. However, in those jurisdictions that have retained parole and indeterminate sentencing, the precise amount of time a convicted offender actually serves is determined by judgments regarding the inmate’s behavior made by prison officials over the course of his or her confinement. Such judgments in turn influence decisions regarding the classification, discipline, and ultimate release date of the inmate. For this reason, those prison officials with direct day-to-day contact with inmates are in a position of considerable power over the length of the sentence individual inmates will ultimately serve.

In private prisons, it is the employees who exercise this power. From the perspective of the parsimony principle, the concern with this arrangement arises from the possibility that private prison operators, whose profitability depends on maintaining a high occupancy rate, could encourage their employees in subtle and not-so-subtle ways to make judgments regarding individual inmates’ behavior so as to prolong the amount of prison time that inmates serve, regardless of whether such extensions are consistent with the demands of legitimate punishment. True, at least in some states, legislators have sought to mitigate the danger of such manipulation by reserving to state officials the final authority over determinations bearing on length of sentence, including changes to inmate classification and any decisions relating to inmate release dates, parole decisions, work release, and reduction of good-time credit. However, even under such arrangements,
private prison employees still wield considerable influence over the administrative proceedings that affect individual inmates’ length of confinement: as one CCA employee, charged with reviewing disciplinary cases, put it, “I’m the Supreme Court.” The formal reservation of these powers to the state may thus be insufficient to counteract the dangers such influence represents for the legitimacy of criminal punishment in terms of parsimony.

Take, for example, the disciplinary context. Discipline in prisons is kept by guards, who have authority to “write up”—that is, to issue disciplinary tickets (D-tickets) to—inmates caught violating prison regulations. Following the receipt of a D-ticket, the inmate will be called to a hearing (D-hearing), at which time evidence may be entered and testimony heard and after which the hearing officer will issue the verdict and pronounce sentence. Depending on the infraction, penalties may include revocation of good-time credits and thus the extension of the inmate’s term of incarceration.

Because inmates have no right of counsel at D-hearings, it is generally their word against that of the guard who wrote up the infraction. Under such circumstances, even in the public sector, the inmate is at a considerable disadvantage: given the solidarity among corrections officers, which can frequently devolve into a mentality of “us” against “them,” the hearing officer’s sympathies tend to lie with the disciplining officer. When the facility is run by a private, for-profit corporation, the worry is that the process will be skewed even more strongly against the inmate. The guard writing up the infraction, and in many cases the hearing officer as well, will be employed by a corporation with a direct financial stake—indeed, a paramount interest—in maintaining a high occupancy rate. This arrangement raises the concern that official testimony and judgments rendered at D-hearings will not reflect the treatment that the inmates deserve or that is consistent with the state’s interest in imposing only legitimate punishments, but will instead reflect the financial interests of the company running the prison.

The same can be said for parole decisions. In the decision of any parole board, the inmate’s disciplinary record while in prison carries great weight. Indeed, in many cases, prison officials are “called upon to provide parole boards with testimony [and] parole recommendations.” Again, the worry is that private contractors’ financial interest in the outcome of parole proceedings “may impair private officials’ objectivity” in a way that yields parole denials for otherwise qualified inmates.

Little study has been made of this aspect of private prison life, and as a result little definitive proof exists of widespread abuses of discretion of the type just postulated. Given the demands of the integrity condition, however, to raise a salient parsimony concern it is not necessary to have definitive empirical proof that the feared abuses have in fact taken place. It is enough that the policies in question create an appreciable risk that illegitimate interests will affect the nature and scope of punishments imposed by the state.

Here, the Supreme Court’s due process jurisprudence provides an apt analogy. The Court has held that it violates due process to subject the “liberty or property [of any defendant] to the judgment of a court the judge of which has a direct, personal, substantial, pecuniary interest in reaching a conclusion against him in his case.” In Connally v. Georgia, for
example, the Court held that the issuance of a search warrant by a justice of the peace violated due process when the justice was paid a fee of five dollars only when he issued the warrant sought, and received no compensation if the warrant request was denied. Such a situation, the Court found, “offers ‘a possible temptation to the average man as a judge . . . [that] might lead him not to hold the balance nice, clear and true between the State and the accused.’” Indeed, the Court has gone still further and required “that judges avoid even the appearance of financial bias.” As the Court explained, “to perform its high function in the best way, ‘justice must satisfy the appearance of justice.’” These cases are plainly concerned with the substantive fairness of the outcome, the worry being that the decisionmaker will not be truly impartial and will thus reach a questionable result. But they are also concerned with the integrity of the system itself, and with ensuring that those subject to punishments can trust in the impartiality of the decisionmaker and thus in the legitimacy of the verdict. For where the judge is known to have a vested interest in the outcome, the subjects of punishment are unlikely to trust and respect the judgment issued.

This same set of concerns applies to the private prison context. True, private prison guards’ financial incentive to maintain high occupancy rates is not as stark as the justice of the peace’s financial incentive in Connally, because private prison employees, unlike the justice of the peace, receive a salary regardless of any role they might play in disciplinary or parole hearings. But the interest of their employers in this regard is still substantial; depending on the contract price, a single parole denial or revocation of good-time credit could be worth as much as $10,000 to $20,000 or more per year, a notable sum in an industry where contractors work on extremely narrow profit margins. The guards, who are hired to work in the interest of their employers and whose job security is dependent on the financial success of the operation, surely know that this is so. Under such circumstances, it would come as no surprise if illegitimate decisions were in fact rendered by private prison employees, or if prison inmates, along with their families, friends, and communities, were skeptical of the impartiality of the decisionmakers and thus unable to trust the legitimacy of judgments affecting the length of their incarceration.

In public prisons, too, the impartiality of corrections officials can be infected by illegitimate motives, by a distaste for particular inmates, or by sheer delight in exercising power. All such motives could—and likely do—lead public correctional officers to issue D-tickets, to revoke good-time credits, or to testify at parole hearings in ways at odds with the parsimony principle, even absent the financial interest of private prison providers. However, this likelihood does not vindicate private prisons. Rather, recognizing the particular danger the profit motive creates in the context of private prison discipline and parole prompts consideration of other, less obvious ways that dispositions in prison disciplinary and parole hearings may be illegitimately skewed.

C. Influencing Incarceration Rates from the Outside: “The Most Powerful Lobby You’ve Never Heard Of”

The private prison industry, to increase the demand for its services, exerts whatever pressure it can to encourage state legislators to privatize state prisons. This effort does not necessarily suggest a parsimony concern, for the fact of privatization alone need not
affect the number of individuals who are actually incarcerated or the length of prison sentences. But what if the private prison industry were exerting political pressure on state legislators not only to encourage a shift to privatization, but also to generate harsher sentencing regimes? This would create the possibility that the state’s sentencing policies, and thus the sentences imposed pursuant to them, are inconsistent with the priority of the most urgent interests and instead serve the financial interests of the private prison industry and the politicians who accept campaign contributions from industry members. By creating an industry capable of, and with an interest in, corrupting the legislative conditions for legitimate punishment, the state’s use of private prisons would be directly at odds with the demands of the parsimony principle.

Given the financial interest of private prison providers in increased incarceration rates, it would not be unexpected if the industry did seek to influence legislation in this direction. Interestingly, however, although the industry is adept at lobbying legislators and targeting campaign contributions to promote its privatization agenda, there is little evidence of any such efforts in support of harsher criminal sentencing schemes. Some commentators, noting this fact, have argued that the private prison industry has no need to push for stiffer sentences. They suggest that by the time the industry emerged in the mid-to-late 1980s, the prisons were filling up so quickly as a result of other, unrelated forces that prison contractors have not needed to undertake any deliberate efforts to ensure continued demand for their services.

However, even if demand for prison space is currently sufficient to ensure the financial position of private prison operators, there is no guarantee that the same will be true in the future. Here, the analogy of the U.S. defense industry, suggested by J. Robert Lilly and Matthew Deflem, may be instructive. After the Cold War, military contractors in the United States suffered a decline in the demand for their services. In response, explain Lilly and Deflem, members of this group “successfully lobbied for governmental concessions and support in the form of changes in the guidelines for selling arms to foreign customers.” Previously, military contractors only received U.S. government approval for foreign arms sales when the sales were found by state officials to “support[] American foreign policy goals and strengthen[] regional alliances.” However, after successful lobbying by the industry hoping to expand its markets in a period of declining American investment in defense, the guidelines for such sales were amended, and they now “require that the U.S. government also consider their benefits to the nation’s military contractors.”

In other words, it was not until economic opportunities for defense contractors began to shrink that the industry began pressuring legislators to generate policies consistent with its corporate interests. This experience suggests that even if private prison providers have had no need as yet to pressure state legislators to shape sentencing policies consistent with their financial interests, these conditions are subject to change. Ultimately, the worry is that, as in the case of the defense industry, the power, wealth, and political connections of the corrections industry may mean that “concerns for profit, efficiency, competition, and money may radically alter the . . . normative goals [of the corrections domain].”
The defense industry analogy suggests that a lack of evidence as to present efforts to unduly influence sentencing policy does not necessarily mean that the parsimony concern is misplaced in the context of private prisons. It is, moreover, not so clear that the private prison industry has not yet taken steps to promote the adoption of statutes likely to increase the size of the prison population. Although industry members have taken little overt action in this direction, at least two private prison firms—industry leaders CCA and Wackenhut—have for some time been “private sector members” of a little-known organization called the American Legislative Exchange Council (ALEC). ALEC is an unusual organization. It takes no public credit for its legislative successes, instead preferring to maintain a low profile. Its main function is the drafting of model legislation, which its legislator-members take back to their home jurisdictions and do their best to turn into law. In 2000 alone, over 3,100 bills based on ALEC’s model legislation were introduced into legislatures by its members, with 450 such bills signed into law. This success is a function of the sheer number of legislators from around the country who are members of ALEC—2,400, almost a third of all state and federal legislators nationwide.

With minimum annual dues of $5,000, corporations and trade associations can also become ALEC members. Upon further payment of “applicable Task Force Dues” (ranging from $1,500 to $5,000 annually), private-sector members can join the task force of their choice and participate in drafting the model legislation that public-sector members will introduce to their respective legislatures.

According to ALEC itself, the criminal justice task force has been among the organization’s most successful working groups. Its successes have included the passage in forty states of ALEC-sponsored “truth in sentencing” statutes, mandating that convicted offenders serve at least 85 percent of their sentences before being eligible for parole, and the passage in at least eleven states of three-strikes laws, which impose mandatory minimum sentences of anywhere from twenty-five years to life for offenders convicted of a third serious offense. Although ALEC takes care to obscure the role played by corporations standing to benefit from its legislative initiatives, it is known that both CCA and Wackenhut Corrections have been private-sector members of ALEC, that both have been among its “major benefactors,” and that at least CCA has participated in the drafting sessions of ALEC’s criminal justice task force, including, reportedly, that session which produced ALEC’s model truth-in-sentencing bill. Indeed, this task force has been cochaired by representatives from CCA at least twice, once by John Rees, a CCA vice president, and once by Brad Wiggins, CCA’s director of business development.

It is impossible to know what direct role, if any, representatives of CCA and/or Wackenhut have played in the drafting and promotion of ALEC-sponsored legislation likely to expand the prison population. It is clear, however, that each company pays thousands of dollars in annual membership dues for a seat at the drafting table with influential legislators. They do so, furthermore, under the auspices of an organization committed to policies certain to increase prison populations nationwide in a way that is consistent with the contractors’ own financial interests. This fact at the very least makes skepticism appropriate regarding the claim that private prison providers take no part in promoting legislation that puts upward pressure on incarceration rates. Moreover, it
reveals that, should these leading members of the private prison industry see a financial benefit to themselves in promoting harsher sentencing regimes, there are channels through which they might effectively further these interests—channels that are conveniently out of public view. Given the distaste with which the public might well react to the notion of private prison contractors lobbying for stiffer criminal penalties—a move that would suggest a cynical willingness to lock more people up for longer periods so that CCA and its fellow industry members might profit financially—their involvement in ALEC may signal, not a lack of interest in promoting such legislation, but a recognition that such efforts might best be undertaken behind closed doors.

The link between CCA, Wackenhut, and ALEC’s criminal justice initiatives provides no concrete evidence of undue influence over sentencing policy on the part of private prison providers. It nonetheless effectively illustrates the tension between the state’s use of private prisons and the demands of the parsimony principle’s integrity condition. At the heart of this condition is the imperative that the state do all it can to secure the conditions of legitimate punishment and to avoid taking steps likely to corrode these conditions. Yet the state, through the use of private prisons, not only allows but facilitates the growth of an industry with a “direct, personal, substantial, pecuniary interest” in increased incarceration rates regardless of the demands of legitimate punishment, with no particular commitment to ensuring legitimate punishment, and with direct access to powerful legislators through both public and back channels. Viewed in this light, it is hard not to see the state’s support of the private prison industry as inviting the possibility that this constituency will exercise undue influence on sentencing policies. Under these circumstances, especially given the extent of the campaign donations given by the private prison industry to key legislative players, citizens may well wonder about the legitimacy of sentencing policies and the punishments imposed pursuant to them. And this suspicion is likely to be strongest in communities most vulnerable to state punishment of any sort—those communities that themselves enjoy little political access or influence.

D. CCA & CCPOA: “Protecting the Public Interest”?

There is thus a tension between the existence of a successful and influential private prisons industry and the demands of the parsimony principle. But private prisons are by no means unique in the threat they pose to the legitimacy of particular punishments and to citizens’ trust in the institutions of the criminal justice system. This threat exists whenever sentencing policy is influenced by interest groups with a strong financial interest in increased incarceration and longer prison sentences. And were private prison providers to seek to wield such influence themselves, they would enter a politicized arena in which several other interest groups already work to shape criminal justice policies in ways consistent with the financial interests of their members.

Perhaps the most notable example in this regard is the California Correctional and Peace Officers’ Association (CCPOA). This organization, one of the most powerful lobby groups in California, represents all of California’s correctional officers and consistently supports state legislation providing for enhanced sentencing, seemingly regardless of the legitimacy of the punishments thereby imposed. The existence of CCPOA and other criminal-justice interest groups, however, does not vindicate the state’s use of private
prisons, as some commentators appear to believe. Any time criminal justice policy is influenced by parties hoping to further their financial interests through increased incarceration regardless of the demands of legitimate punishment, it is cause for concern. The fact that the private prison industry is not the only group motivated in this direction suggests not that there is no problem with private prisons, but that the problem is more widespread than previously recognized.

As private prison advocate Charles Logan sees it, introducing the private prisons industry into the political mix better serves “the public interest” by forcing competing interest groups—correctional officers’ unions, state agencies, etc.—to press their claims in the most vociferous way possible. This process, Logan claims, allows policymakers and citizens to “sort[] out the [public interest] from among competing definitions and claims.” My own view is somewhat different. To satisfy the parsimony principle, the state is obliged to avoid taking steps likely to corrupt the conditions of legitimate punishment. Yet the presence of any powerful interest group with a financial stake in increased incarceration creates the danger of such a corrupting influence. Granted, this sort of political pressure is routinely brought to bear by interest groups of all kinds hoping to influence all manner of legislation. But whatever one might think of this system more generally, it is out of place when the issue is criminal punishment.

In the case of punishment, the state is taking the extraordinary step of heavily burdening the security and integrity of individual citizens. Imposing such a burden is not beyond the scope of the state’s legitimate power. But if the exercise of this power is to be legitimate, it must be consistent with the priority of the most urgent interests. And if the criminal justice system is to earn citizens’ trust, the process for setting the terms of state punishment must be driven by a good-faith effort by all parties to craft policies consistent with the demands of legitimate punishment. However, where it is known that the process and state officials themselves are open to the influence of parties standing to benefit financially from increased incarceration, not only may the punishments flowing from particular criminal justice policies prove illegitimate, but citizens are also likely—rightly—to view them as such.

The private prison industry and the correctional officers’ unions are not the only entities with a financial interest in increased incarceration rates. As Lilly and Knepper have shown, “there are many more companies profiting from the routine, low-profile world of providing prison services,” among them private companies who provide prisons with services including “food service design and management, consulting and personnel management, architecture and facilities design, vocational assessment, medical services, drug detection, and transportation. . . . And that is not all.” There are also those companies that supply prisons and jails with equipment, selling them, among other things, “protective vests for guards, closed-circuit television systems, mechanical and electronic locks, perimeter security and motion detection systems, fencing, flame-retardant bedding, furniture, foot [wear], lighting, and linen along with shatter-proof plastic panels, tamper-proof fasteners, and clog-proof waste disposal systems.” As this account suggests, prisons are big business. There is thus a wide range of private interests in a position to profit from an increase in the number of people incarcerated, and potentially a large number of interest groups with the desire, and perhaps the financial
wherewithal, to seek to influence sentencing policies in ways consistent with their financial interests.

Widening the lens to include this range of parties with a financial interest in incarceration may seem to extend the parsimony concern too far, perhaps thereby negating its critical bite. Am I saying that no one in contemporary society should be able to make money from operating prisons? If so, what about correctional officers, who work for salaries and benefits? And how would prisons receive the goods and services that are incontestably necessary if they are to run at all? The point here is not that, for punishment to be legitimate, no one can benefit financially from corrections; as the questions just posed indicate, such a requirement would be a practical impossibility. That this is so, however, does not require averting one’s eyes from the potential dangers created when people or entities with a strong financial interest in increased incarceration are also positioned to influence the nature or extent of punishments imposed. Where such circumstances exist, it is essential to call attention to them and be explicit about the threat they represent. Society must also do all it can to protect the process of crafting sentencing policies from any undue influence.

Admittedly, short of generating an explicit and widespread commitment among legislators and their constituents to satisfying the conditions of legitimate punishment, a total abatement of the sort of undue influence over sentencing policy described here is unlikely. Absent that commitment, the interest-group model of politics that reigns in the criminal justice context will not likely be widely condemned or even questioned. Campaign finance reforms designed to constrain illegitimate influence would certainly be welcome, as would tighter conflict-of-interest rules, for example, those prohibiting legislators or their spouses from owning stock in companies that stand to gain significant economic benefits from increased incarceration. But any such efforts are likely to be limited in their effects, for several reasons. First, even absent overt lobbying and obvious conflicts of interest, there is reason to expect a sympathy of perspective and priorities between policymakers and at least some private interests. This is particularly so in the private prison industry, where the “revolving door” between government and private corrections firms means that private prison executives will often have a considerable range of contacts in the government of the state whose prison markets they seek to access. Second, effective lobby groups will still be able to find mechanisms to promote their financial interests in ways less obvious—and thus less open to regulation and less susceptible to critique—than direct lobbying of legislators. And third, these methods will do little or nothing to prevent the kind of access and influence enjoyed by private-sector members of organizations like ALEC. Yet even granting that campaign finance reform efforts are likely to fall short, society remains obligated to recognize the broader legitimacy problem posed by the possibility of private influence over the legislative process and to do what it can to ameliorate it.

E. Prison Building as Economic Development: “Recession-Proof Jobs”

There is a further reason why the concerns raised by the parsimony principle cannot be fully addressed through initiatives targeted at constraining the undue influence of identifiable interest groups over the legislative process: the tendency to look to prisons—
and prisoners—as a revenue source is not restricted to interest groups like these. In addition to the groups already discussed, communities across the country, led by public officials at all levels, have come to view prison building as the means to bring in jobs and grow local economies. As a result, voters and their political representatives now regard incarceration as a means to promote their own financial interests. Thus these groups, too, may well have an incentive to support policies likely to increase the prison population regardless of the implications for legitimate punishment.

Take, for example, the case of Wise County, Virginia. In the late 1990s, with much fanfare, the state of Virginia opened two state-of-the-art supermax prisons, Wallens Ridge and Red Onion, in Wise County. To do so was a serious investment; Wallens Ridge alone cost the state $77.5 million, an average of $110,085 per cell.

The supermax prison is a recent innovation in incarceration. As the term “supermax” suggests, the purpose of these prisons is generally to house what prison administrators refer to as “the worst of the worst”—those inmates who are too violent or unruly to be kept under control even in the relatively restrictive conditions of conventional maximum security. Originally, the intended purpose of Virginia’s new supermax facilities was to house this subset of prisoners. Although “never precisely defined,” the worst of the worst was understood by the state’s director of prisons to “include[] two categories of inmates: those who had been ‘disruptive’ at other prisons (those who had attacked other inmates or guards) and those who had been sentenced to terms of life or ‘near life’—typically, eighty years or more.” As it happened, however, between Wallens Ridge and Red Onion, the Virginia Department of Corrections (DOC) had more supermax cells than qualified prisoners. In response, the DOC “quietly expanded the eligibility” for supermax classification. By the time both prisons were opened, they were accepting inmates sentenced to as few as five years, who had harmed no one and had never been disruptive.

Incarceration in supermax represents a qualitatively different and more severe punishment than even incarceration in a conventional maximum-security prison. Inmates housed in such prisons are severely restricted in their movements and actions. They are generally locked down in their cells twenty-three hours a day, with one hour given to (solitary) physical activity in an outdoor “concrete exercise ‘yard’ that is simply a larger version of their own cell, minus the toilet and roof.” There is neither programming nor inmate interaction, and possession of personal items is extremely limited. Why would the state of Virginia extend its control in this way over inmates whose behavior did not warrant it? The reason is economic development. In 1995, Wise County had seen the closing of the Westmoreland Coal Company, then the biggest employer in the area, and the building of the prisons was intended to bring business and jobs to the area. As Governor James Gilmore said at the opening of Wallens Ridge, the prison “is an economic boon for this town and this county and this region . . . [a] win-win for everyone.” A prison operating at less than full capacity, however, means less of an economic boon than otherwise possible: fewer jobs for local residents, fewer visitors to the area, less business for area merchants, and even fewer mouths to feed inside the prison walls. Hence the expanded supermax eligibility.
Wise County is just one example of a community looking to prisons—and the imposition of punishment they represent—for economic development purposes. But such examples abound. Throughout the United States, communities in rural areas in particular have come to look to prison building as the way to boost their economies, provide jobs, and increase tax revenues, and even to increase the local population registered by the national census, thereby increasing the state and federal subsidies for which the community is eligible. In the 1990s, an estimated 245 prisons were built in “212 of the nation’s 2,290 rural counties, many in Great Plains towns of Colorado, Oklahoma and Texas that had been stripped of family farms and upended by the collapse of the 1980s oil boom.” These communities see prison building as a safe, “clean” form of economic development, and they also see it as “recession-proof”; as one city manager commented, “Nothing’s going to stop crime.” State legislators appear to share this view, and routinely vie to locate new prisons in their districts to provide an economic boon for constituents.

The phenomenon of what one observer has called “prison construction advocacy” does not necessarily mean that the punishments served in the prisons built under these circumstances are illegitimate. Viewed from the perspective of the parsimony principle, however, the worry is that maintaining a high incarceration rate will become fused in the minds of voters—and, even more troubling, in the minds of their political representatives—with the possibility of economic prosperity, potentially corrupting the process of determining appropriate criminal punishments with concerns that are plainly illegitimate. Interestingly, in many cases, prison building has not provided the economic security and employment opportunities for residents that communities often expect. But the federal and state financing a county receives is based on census data, which includes the inmate population of local prisons. Having a prison in the district can thus bring a community a considerable amount of extra money, whether or not the employment rate is positively affected. And even in cases where building a prison brings fewer economic benefits than might have been hoped, once the prison is built and filled, citizens still have a strong financial incentive to see incarceration levels maintained.

Again, I know of no empirical evidence confirming that the allure of economic benefits directly inclines voters or their representatives to push for “tough-on-crime” initiatives. From the perspective of the parsimony principle, however, it is enough that the incentives to do so exist, promoted by the very legislators charged with the task of defining the terms of legitimate punishment for convicted offenders. Such incentives create the real possibility that illegitimate interests will affect the nature and extent of punishments ultimately imposed. They also give citizens—especially those citizens most likely to be subject to criminal punishment—reason to mistrust the process, grounding the suspicion that criminal punishments are imposed not for legitimate reasons but instead to sustain the economies of voters in rural communities and the reelection prospects of their political representatives.

If this concern seems far fetched, consider the racial dynamics that arise from treating prisons as engines of rural economic development. The communities that have benefited from this strategy are overwhelmingly “white, rural, Republican enclaves.” The prisoners shipped in to populate the new prisons, in contrast, are overwhelmingly “minorities from urban, Democratic areas.” Many of these prisoners, moreover, are serving time for
nonviolent offenses. Under these circumstances, it is not unreasonable to think that, when prisoners are viewed by local whites as the key to prosperity, those same prisoners—not to mention their families, friends, and communities—would come to view their incarceration as more about jobs and revenue for rural communities than about satisfying the demands of legitimate punishment.

F. Private Prisons as Miner’s Canary

Considering the state’s use of private prisons through the lens of the parsimony principle reveals a possible threat to the legitimacy of punishment whenever parties with a financial interest in increased incarceration are in a position to exert influence over the nature and extent of criminal sentencing. If this concern is real, it suggests an independent reason for the state not to privatize its prisons: even granting that similar concerns exist elsewhere in the criminal justice system, the state ought not to foster yet another potentially influential industry that could seek to compromise further the possibility of legitimate punishment to promote that industry’s own financial interests. This is especially so given the likely limitations of available mechanisms for constraining any undue influence by private parties on criminal justice policies.

But exploring this concern in the context of private prisons has revealed a problem that extends well beyond this context, arguably reaching the core of our majoritarian system. That this is so vindicates the assertion, made above, that considering private prisons from the perspective of liberal legitimacy provides a lens through which to see in a new light practices in the criminal justice system that may have been too readily taken for granted. At the same time, however, the apparent extent of the problem may seem to undercut the value of the exercise; having revealed so deep a problem, I can offer no easy remedy.

The difficulty is that, absent widespread commitment to ensuring that criminal punishment satisfies the demands of the parsimony principle, there will be no broad sympathy for the view that society needs to exclude illegitimate considerations like the financial advantage of voters from the process of establishing sentencing policies. This means that to address the problem properly, there can be no half measures. So long as one continues to assume that criminal justice policy is appropriately shaped through an interest group model of politics, where all parties with a stake in the outcome vie for policies friendly to their own interests, looking to prisons for economic prosperity will seem entirely unobjectionable. Indeed, on the interest-group model of politics, it seems exactly right. The only adequate remedy is broad acceptance of the idea that the parties charged with determining the nature and extent of criminal punishment have an obligation to make such determinations in isolation from any possible knowledge of their own personal interests—or those of their constituents.

Broad acceptance of this approach, however, is a tall order; it would, as Simone Chambers recently noted, require that legitimate punishment come “to be understood democratically, not just philosophically.” Certainly, any such process of “public opinion formation” will not happen overnight. But the first step is to make explicit the potential conflict between legitimate punishment and conceiving of incarceration as a source of financial gain. Having done so, it will then be possible to challenge state officials to act in
ways that minimize this conflict as much as possible, and to do all they can to rise above this impulse themselves.

In the private prisons literature, it has largely been assumed that prisons are no different from any other government function to be privatized. Private prisons have thus been treated as an issue of privatization and not one of criminal justice policy. The foregoing analysis, however, makes plain what is lost with this move: it makes it impossible to see that the state’s use of private prisons reflects a larger trend toward viewing incarceration in economic terms and regarding prison inmates as the economic units of a financial plan. If anything, private prisons appear to be the logical extension of this vision, which already informs myriad aspects of this country’s criminal justice system, including the practice of prison administrators contracting out the provision of basic services to cut the cost of corrections; underinvestment in mechanisms for accountability and oversight; and the tendency of private prison providers, correctional officers, and the voters themselves to look to increased incarceration as the means to their financial well-being.

From the perspective of liberal legitimacy, there is a serious cost to the widespread adoption of this economistic view: society becomes less likely to see those it punishes as human beings and more likely to lose a sense of the severity of the burdens punishment imposes. Indeed, from the perspective of liberal legitimacy, the most troubling thing about private prisons may be what they reveal of this country’s collective failure of respect and responsibility toward those it incarcerates.

CONCLUSION

The debate over private prisons has largely been framed as a choice between alternative management structures judged on the basis of their relative efficiency. But framing the issue in terms of comparative efficiency is the wrong way to think about private prisons, for at least two reasons. First, this approach leads to an undue focus on the value of efficiency, when what should be of most concern in the incarceration context is whether the penal practices at issue are in fact legitimate. Second, comparative efficiency uncritically accepts the current state of prison conditions as the appropriate baseline for analysis, judging private prisons to be good enough when conditions of confinement within them meet the standard set by public prisons, thus providing no opening for challenging this (low) baseline as unacceptable.

Focusing on the comparative question, that of whether the management structure of prisons should be public or private, is to lose sight of the bigger picture. The real question is why all prisons, public and private alike, fall so far short of satisfying society’s obligations to those it incarcerates. Exploring the state’s use of private prisons from the perspective of liberal legitimacy has helped to answer this question by making plain what is wrong with private prisons in the form currently on offer, and, in so doing, exposing as problematic several practices operating within the penal system as a whole which tend to be taken for granted and thus no longer questioned. The foregoing analysis has, for example, exposed the danger of delegating to prison officials extensive discretion and power over a largely vulnerable and dependent prison population in the absence of adequate accountability mechanisms—and demonstrated that this danger is particularly
acute when those officials are motivated by private purposes at odds with the possibility of humane punishment. It has revealed the threat posed to humane conditions of confinement when state officials seek to save money by contracting with for-profit entities for the provision of crucial prison services. And it has indicated the corrosive effect, both on the possibility of legitimate punishment and on citizens’ trust in the criminal justice system itself, of the unquestioned idea that sentencing policy is appropriately influenced by advocacy groups with a financial interest in increased incarceration. The examination of private prisons from the perspective of liberal legitimacy suggests that a meaningful commitment to the possibility of legitimate punishment requires that all these practices be curtailed, or, at the very least, engaged in warily, not only in the private prisons context, but in the context of the penal system in general.

As noted in the opening pages of this Article, a recurring theme among opponents of private prisons is that incarceration is an inherently public function, one that cannot be legitimately delegated to private actors. But, as the foregoing discussion has revealed, the fact that punishments of incarceration are served in prisons administered by state employees is no guarantee of their legitimacy. To the contrary, public prisons too can be sites of unchecked discretion exercised by individuals with their own personal interests and agendas, whether those individuals are state-employed corrections officials or private contractors engaged by public prison officials to provide discrete prison services as a way to save states money. And the punishments served in public prisons are as likely as punishments served in private prisons to be shaped by the efforts of political actors seeking to further their own financial interests.

Still, there is arguably something to the view that punishment, if it is to be legitimate, should be a wholly public function, untainted by private motives and interests. But if so, it is not just the use of private prisons, but the current approach to criminal punishment in general—shot through as it is with scope for furthering private purposes at the expense of convicted offenders—that sorely needs to be rethought.


A common argument against privatization is that private providers will self-interestedly lobby to increase the size of their market. In this Article, I evaluate this argument, using, as a case study, the argument against prison privatization based on the possibility that the private prison industry will distort the criminal law by advocating for incarceration.

I conclude that there is at present no particular reason to credit this argument. Even without privatization, actors in the public sector already lobby for changes in substantive law—in the prison context, for example, public corrections officer unions are active advocates of pro-incarceration policy. Against this background, adding the “extra voice” of the private sector will not necessarily increase either
the amount of industry-increasing advocacy or its effectiveness. In fact, privatization may well reduce the industry’s political power: Because advocacy is a “public good” for the industry, as the number of independent actors increases, the dominant actor’s advocacy can decrease (since it no longer captures the full benefit of its advocacy) and the other actors may free ride off the dominant actor’s contribution. Under some plausible assumptions, therefore, privatization may actually decrease advocacy. Under different plausible assumptions, the net effect of privatization on advocacy is ambiguous, but in any event, privatization does not unambiguously increase advocacy.

The argument that privatization distorts policy by encouraging lobbying is thus unconvincing without a fuller explanation of the mechanics of advocacy.

INTRODUCTION

Over ninety years ago, opponents of World War I alleged that “munitions manufacturers frighten the popular mind with the fear of imaginary external enemies and inflame it with murderous patriotism.” According to a view attributed to Stefan Zweig, the war began only when “newspapers in the pay of the arms manufacturers began to whip up sentiment against Serbia.” After the war, that accusation morphed into the charge that arms makers were self-interestedly obstructing peace efforts. Today, an opponent of U.S. military policy characterizes defense contractor CACI International, Inc., whose chairman speaks publicly of the “heinous[ness],” “fanatical horror,” and “barbarism” of terrorism, as “one of the most unabashed corporate backers of Bush’s foreign policy and a key supporter of the military campaigns in Iraq and Afghanistan.” Critics also charge that private military interests affect what weapons systems we rely on and what alliances we enter into, and that, in some countries, those interests may even take over the government.

This theme—that private contractors use their influence to advocate not just more privatization but also, insidiously, changes in substantive policy—sweeps more broadly than just defense contractors. The following list gives a sense of the generality of the accusation; the last few items illustrate that the critique comes from “the right” as well as from “the left.”

- Private prison firms are often accused of lobbying for incarceration because, like a hotel, they have “a strong economic incentive to book every available room and encourage every guest to stay as long as possible.”

- Business improvement districts—coalitions of business and property owners, many of which have their own private security forces—have lobbied municipalities for, among other things, aggressive panhandling ordinances.

- A toll road developer in Colorado has lobbied for statutory changes to preempt county authority to set toll rates, and a private road construction firm has been accused of contributing to Texas Supreme Court justices’ campaign chests to influence a potential eminent domain suit related to a toll road in the state.
Private landfill companies have been accused of lobbying for weak environmental regulation of landfills and opposing recycling initiatives.

Private water-supply owners have been accused of “lobbying to weaken water quality standards . . . and pushing for [trade agreements] that hand over the U.S. water resources to foreign corporations,” and private water utilities have been accused of fighting conservation efforts.

Private redevelopment corporations, which have the power to condemn private property for purposes of “urban renewal,” have opposed reform of eminent domain laws in the wake of the Supreme Court’s decision in *Kelo v. City of New London*.

And “private attorneys general,” for instance environmental groups that benefit from fines available under environmental citizen suit provisions, or members of the securities plaintiffs’ bar who benefit from the availability of securities fraud class actions, fight for the continued vitality or even strengthening of the statutes under which they litigate.

In this Article, I examine this “political influence” challenge to privatization using the case study of private prisons. I conclude that, in the prison context, there is at present no reason to credit the argument. At worst, the political influence argument is exactly backwards, by which I mean that privatization will in fact *decrease* prison providers’ pro-incarceration influence; at best, the argument is dubious, by which I mean that its accuracy depends on facts that proponents of the argument have not developed.

Private prisons are a useful case study. First, they are a growth industry, having progressed from humble beginnings in the late seventies and early eighties to now house about one in sixteen prison inmates nationwide. Second, the opponents of private prisons commonly make the political influence argument.

For example, in a recent *Duke Law Journal* article, Sharon Dolovich writes that “the legitimacy of punishment” is threatened “whenever parties with a financial interest in increased incarceration are in a position to exert influence over the nature and extent of criminal sentencing. If this concern is real”—and she suggests that it may well be—prisons should not be privatized because “the state ought not to foster yet another potentially influential industry that could seek to compromise further the possibility of legitimate punishment to promote that industry’s own financial interests.”

David Shichor, a prominent contributor to the prison privatization literature, opposes prison privatization in part because:

> Through political lobbying, PACs, campaign contributions, and the provision of perks to politicians (as industrial and business corporations do), corporations are likely to continue to support and even accelerate incapacitation-oriented legislation and policies by which more people will spend longer periods of time in correctional institutions. Conversely, this trend may diminish the emphasis on
alternative programs and will result in the pursuance of the “Hilton Inn mentality,” that is, trying to maintain high occupancy rates for profit purposes.

And Brigette Sarabi and Edwin Bender’s thesis is clear from the title of their report, *The Prison Payoff: The Role of Politics and Private Prisons in the Incarceration Boom*, in which they argue that prison privatization should be resisted in part because private prison firms have a “vested financial interest[] in increasing rates of imprisonment.” This is only a small sample of the literature. . . .

I assume, for purposes of this Article, that the concern underlying this critique is reasonable—that is, that economically self-interested pro-incarceration advocacy is undesirable. This concern, however, fails to support the argument against privatization for several reasons.

First, self-interested pro-incarceration advocacy is already common in the public sector—chiefly from public-sector corrections officers unions. For instance, the most active corrections officers union, the California Correctional Peace Officers Association, has contributed massively in support of tough-on-crime positions on voter initiatives and has given money to crime victims’ groups, and public corrections officers unions in other states have endorsed candidates for their tough-on-crime positions. Private firms would thus enter, and partly displace some of the actors in, a heavily populated field.

Second, there is little reason to believe that increasing privatization would increase the amount of self-interested pro-incarceration advocacy. In fact, it is even possible that increasing privatization would *reduce* such advocacy. The intuition for this perhaps surprising result comes from the economic theory of public goods and collective action.

The political benefits that flow from prison providers’ pro-incarceration advocacy are what economists call a “public good,” because any prison provider’s advocacy, to the extent it is effective, helps every other prison provider. (We call it a public good even if it is *bad* for the *public*: the relevant “public” here is the universe of prison providers.) When individual actors capture less of the benefit of their expenditures on a public good, they spend less on that good; and the “smaller” actors, who benefit less from the public good, free ride off the expenditures of the “largest” actor.

In today’s world, the largest actor—that is, the actor that profits the most from the system—tends to be the public-sector union, since the public sector still provides the lion’s share of prison services, and public-sector corrections officers benefit from wages significantly higher than their private-sector counterparts’. The smaller actor is the private prison industry, which not only has a smaller proportion of the industry but also does not make particularly high profits.

By breaking up the government’s monopoly of prison provision and awarding part of the industry to private firms, therefore, privatization can reduce the industry’s advocacy by introducing a collective action problem. The public-sector unions will spend less because under privatization they experience less of the benefit of their advocacy, while the private firms will tend to free ride off the public sector’s advocacy. This collective action
problem is fortunate for the critics of pro-incarceration advocacy—a happy, usually unintended side effect of privatization. One might even say that prison providers under privatization are led by an invisible hand to promote an end which was no part of their intention.

This is the simplest form of the story, but one can also tell more complicated versions in which privatization does not necessarily decrease total industry-expanding political advocacy. After presenting my main model, I introduce a few realistic complications. I explain why I am focusing only on public-sector unions and private firms, and whether the individual private firms and the public sector compete or cooperate with each other on advocacy. I alter the assumption that money merely buys the passage of a pro-incarceration measure, and allow money to change the substance of the measure itself. I also relax the assumption that anti-incarceration advocacy is fixed. These complications do not change the basic result of the model. Other complications are more fundamental, and make the effect of privatization ambiguous—increasing private-sector advocacy but also decreasing public-sector advocacy. These complications include relaxing the assumption that the effectiveness of advocacy only depends on the total amount of money spent, and relaxing the assumption that the introduction of privatization into a state is exogenous. If those extensions of the model are closer to the truth, then total advocacy may rise—but it may also fall, depending on which effect dominates. We cannot determine the net effect a priori.

There is thus no reason to believe an argument against prison privatization based on the possibility of self-interested pro-incarceration advocacy—unless the argument takes a position on how lobbying, political contributions, and advocacy work, and why (for instance) any increase in private-sector advocacy would outweigh the decrease in public-sector advocacy. Either this argument against prison privatization is clearly false, or it is only true under certain conditions that the critics of privatization have not shown exist.

The analysis here not only sheds light on the prison privatization debate but also provides a roadmap for analyzing military contracting and other privatization contexts. Because privatization can affect the incentives of both the private and public sectors to wield political influence, one should not conclude that privatization distorts substantive policy in an undesirable direction unless one can tell a story, based on a plausible view of government agents’ behavior, in which private-sector advocacy rises more than public-sector advocacy falls. In the end, each industry has its own idiosyncrasies, so I do not make a strong claim about the use of the argument outside of the prison context. But, at the very least, the use of the political influence argument is often theoretically unsound to the extent it ignores this comparative analysis.

Part I sets forth the main model of the paper. In this model, the sector with the greatest benefit from the expansion of the industry does all the advocacy, and the sector with the smaller benefit entirely free rides off the larger one. Accordingly, privatization reduces industry-expanding advocacy if, after privatization, the public sector remains the sector with the greatest benefit. Part II applies this theoretical model to prisons and suggests, based on an informal calculation, that the actors that would benefit the most from increased incarceration are indeed the public-sector corrections officers unions. Thus, we
should expect the public sector to do all the pro-incarceration lobbying (though less than it would have done without privatization). That Part argues that the simple model, despite its stark result, may be quite close to the truth, as there is a wealth of evidence that public corrections officers unions advocate incarceration, and no such hard evidence on the private side. Part III elaborates on the model, explaining why it is appropriate to focus on public corrections officers unions and private prison firms as the relevant actors, and how cooperation within the prison industry affects the results. Part IV complicates the model in various ways. Some of these complications do not change the basic result of the simple model. Other complications make the result muddier, so that instead of unambiguously reducing advocacy, privatization has a theoretically ambiguous effect on the amount of industry-expanding advocacy.

I. ADVOCACY AS A PUBLIC GOOD

In this Part, I present the main model I use to predict how industry actors will react to privatization. The central feature of the model is that industry-increasing advocacy is a public good. Privatizing part of the industry therefore introduces a collective action problem: unless everyone in the industry cooperates with each other, they will in aggregate spend less on industry-increasing advocacy than a single firm would if it covered the whole industry, because a portion of their expenditures will benefit their competitors.

This intuition should not be surprising, as it is standard in the literature on public goods. When a good is private, everyone pays for, and enjoys, only his own consumption. By contrast, when a good is public, in the classic model, everyone benefits from the total amount, and this amount is determined by the total amount of contribution.

For example, if we benefit from our national defense, we benefit from the full amount, not just from the chunk we paid for; we cannot be excluded from the full benefit, no matter how little we paid; and the total amount of national defense is just determined by how much money Congress allocated to national defense from the Treasury. A tax-funded program that improves air quality benefits everyone who breathes the relevant air, whether or not they contributed to the program, and the total improvement is just determined by the amount of resources directed toward that goal.

Similarly, contributing to a candidate’s campaign benefits all of his supporters, and it is not too implausible to say, as an approximation, that to the extent the money he raises and spends affects his probability of winning, it is only the total amount of money that matters.

In all these cases, the temptation to free ride off one’s peers’ contributions is strong. This Part illustrates the phenomenon of free riding in the context of political contributions.

A. The Basic Model

A monopolist is willing to invest some amount of money in lobbying to increase the size of his industry. To determine that amount, he weighs the benefit that his money can
“buy”—the expansion of the industry is worth something to him, and money can help his policy pass—against the cost of the lobbying.

If that firm is broken up into two smaller firms—say a 90% incumbent firm and a 10% splinter firm—the larger incumbent is not willing to spend as much as it used to, because the costs of lobbying are the same while the benefits are 10% less than they used to be. And the smaller splinter firm will not be willing to spend anything, because it will be satisfied free-riding off the larger incumbent’s lobbying. Thus, splitting up an industry can decrease total industry-expanding lobbying.

The rest of this Part illustrates this intuition graphically.

Suppose you are, as economists say, a rational, risk-neutral expected-utility maximizer. One may dispute how much of life this assumption can explain, but on balance it seems to be at least a good starting point for predicting the behavior of business organizations. You are faced with the choice of whether or not to spend a dollar on political advocacy—donating to the campaign of a politician or voter initiative, contributing to your trade association’s lobbying expenses, or running an ad—in favor of some reform that could increase the size of your market. We may assume that this dollar has some influence in the world, whether appropriate or inappropriate—it could corrupt a legislator, raise the chance of his election, contribute to the passage of the initiative, or change popular opinion.

The benefit of this dollar is the value of the increased probability of getting your desired policy change. It is reasonable to think that spending money on advocacy is subject to decreasing marginal returns, so each additional dollar gets you less and less benefit. The cost of a dollar’s worth of advocacy, on the other hand, is $1—and remains $1, no matter how many dollars you spend. As long as the benefit of an advocacy dollar is greater than $1, you continue spending. As soon as that benefit falls to $1, you stop spending. This is your optimal total amount of advocacy spending—say $1 million.

Figure 2 below illustrates the situation. The expected benefit—that is, the probability of success times the benefit—is represented by the curved line below: the more you spend, the greater the probability of success, but the less you get for each extra dollar. Because a probability cannot get any higher than 100%, the curve is bounded above by the dashed line representing the total benefit of the policy. The cost of advocacy is represented by the straight line below: $1 of spending on advocacy costs exactly $1. Your problem is to maximize the vertical distance between the expected benefit curve and the cost line. In Figure 2, the maximum distance occurs at a spending level of $1 million.
Figure 3 is an equivalent way of seeing the same problem.

The curve [in Figure 3] represents the marginal expected benefit—that is, the benefit of an extra dollar of spending, which is equal to the total benefit times the extra probability of success that a dollar buys you. As noted above, the marginal benefit is decreasing. The straight line is the marginal cost of advocacy: an extra dollar of advocacy spending always costs $1. If the marginal expected benefit is above $1, you’re not spending enough; if it is below $1, you should cut back. At a spending level of $1 million, an additional dollar of spending gives you exactly $1 of expected benefit.

Now suppose the Department of Justice’s Antitrust Division comes in and splits you up, so that you now have 90% of the market and are faced with a competitor with the other 10%. Your previous optimal amount of spending, $1 million, is no longer optimal for you: the cost of that last dollar was $1, and while the benefit of the dollar is $1 for the whole industry, you, who now represent only 90% of the industry, only see 90¢ of that benefit. All your benefits are now lower by 10% because you have to share them with
your competitor. For our purposes, the split-up thus has the same effect as a 10% tax on your benefit. Because your spending on advocacy—an investment in the growth of your industry—is only 90% as productive, you do less of it. You start cutting back on your spending, because a dollar saved puts $1 back in your pocket and only reduces your benefits by 90¢. As you cut back more, the benefit of the last dollar rises; you stop cutting back as soon as the benefit of your last dollar to the industry reaches about $1.11 (which is a $1 benefit to you). Your new amount is, say, $900,000.

This new situation is illustrated in Figures 4 and 5.

\[ \text{FIGURE 4} \]

In Figure 4, the top curve is the expected benefit to the whole industry (as before), and the second curve is your expected benefit, newly reduced now that you have only 90% of the industry. The bottom curve is your 10% competitor’s expected benefit. As discussed above, the maximum vertical distance between your curve and the cost line now occurs at $900,000; and say the maximum distance between your competitor’s curve and the cost line occurs at $300,000.

On Figure 5, the equivalent graph that shows marginal quantities instead of total quantities, you want to find the point where the marginal expected benefit to the industry is $1.11. This is equivalent to finding the point where 90% of the marginal expected benefit (i.e., the benefit to you) is $1. That point again occurs at $900,000. Your competitor wants to find the point where the marginal expected benefit to the whole industry is $10, which gives him $1. This point is at $300,000.
This story is incomplete. You do not want the amount spent to be exactly $900,000; obviously, you would be thrilled if other people happened to contribute more. It’s just that you are not personally willing to put a dollar more into the pot if the pot already contains $900,000. You want the total amount spent to be at least $900,000, and you are willing to contribute money until that point is reached, but you are no longer willing to personally contribute once you are holding the 900,001st dollar. This is because, if the benefit of a dollar only depends on the total amount of money spent, and if the 900,000th dollar had a benefit to the industry worth $1.11 (and thus a benefit to you worth $1), then the 900,001st dollar has a benefit worth slightly less than $1.11.

Your new competitor, who represents the remaining 10% of the industry, and who is equally interested in this reform that will increase the size of the pie, by a similar reasoning, wants the total amount spent to be at least $300,000 and will not put a 300,001st dollar into the pot.

This leads to two conclusions. First, the total amount spent will be exactly equal to the larger actor’s threshold—in this case, $900,000. If it were less, you, the larger actor, would want to spend more money. And if it were more, you would want to take some money out of the pot, since the dollars beyond the 900,000th are giving the industry a benefit below $1.11 and giving you a benefit below $1. Second, there is no reason for your competitor to spend anything. He is unwilling to spend any dollar beyond the 300,000th, since its marginal benefit to the industry is under $10 and its marginal benefit to him is under $1. Thus, suppose you were going to spend $600,000, and he was going to spend $300,000. These would not be equilibrium actions, since he would prefer to keep his $300,000. Why should he spend any extra dollar beyond the $600,000 you are already spending, if the 300,001st dollar already is not worthwhile to him? Thus, the only equilibrium is where you give $900,000 and he gives $0. Because he is the smaller actor, he entirely free rides off you. The result is what Mancur Olson calls the “systematic tendency for ‘exploitation’ of the great by the small.”
B. Industry Shares Versus “Real” Shares

If one accepts the fundamental assumption of this Part—that the probability of success only depends on the total amount of money in the pot—this simple model is flexible enough to accommodate many institutional details of privatization. The total free-riding result happens whenever one actor has a lower threshold than the other, for whatever reason. In this story, you and your competitor are identical except that you have 90% of the industry and he has 10%. But one’s threshold could be lower for other reasons as well.

For instance, suppose that, to add insult to injury, the government not only breaks up your monopoly but also subjects your revenues to a high (50%) tax rate. The breakup already altered your spending threshold by shifting your curves down to 90% of their previous level (compare Figures 2 and 3 with Figures 4 and 5). Now, with the 50% tax, your revenue and marginal revenue curves shift further down—to 45% of their original levels. (If your competitor with a 10% share is subject to the same tax, his curves are 5% of the original industry curves.)

So the combination of the breakup and the tax makes you act like a firm with a 45% market share. These new percentages—call them “real” shares—no longer need to add up to 100% (in fact, with the 50% tax, they add up to 50%), but they convey the economic intuition that your spending threshold is lower when, for whatever reason, your benefits decrease.

After we determine everyone’s “real” shares, the same analysis applies as before: the “biggest” firm does all of the advocacy, and the “smaller” firm is a free rider. The only difference is that we learn who is “biggest” not just by looking at proportions of the market but at shares of total industry revenue. Instead of calling this firm the “biggest” firm, we will call it the “dominant” firm. Thus, if the tax rate on your revenues is 90%, you will act as though your share is not 90% but 9%. If your competitor with a 10% share is exempt from the tax, then he is actually the dominant actor. Now you will free ride off him.

In short, anything that affects your revenues affects your “real” share. Suppose, for instance, that your competitor is less profitable than you are: your 90% share is a monopoly share in 90% of the geographic area, while the remaining 10% is divided among 100 competitors who act according to the textbook perfect competition model, where everyone makes zero economic profits. Then those competitors—and thus that entire 10% of the market—act as though they had a 0% share of the industry.

Or, as a final example, suppose that your competitor is better at advocacy. Perhaps, for whatever reason (maybe he is a slicker lobbyist), each dollar he spends on advocacy is twice as effective as each dollar you spend. Then, he acts as though his share is 20%, and his threshold goes up accordingly. All these considerations affect your “real” shares for purposes of choosing how much to spend on advocacy. (In this example, he still won’t do anything because 20% is still less than your 90% share.)
C. Does Privatization Always Reduce Advocacy in This Model?

This model applies straightforwardly to privatization: partial privatization of an industry splits the industry up into a public sector and a private sector, much as one can split up a monopolistic firm into several competing firms. To be sure, the public sector is not a “profit maximizer” like a private firm. But the concept of profit maximization need not be interpreted in a narrow financial sense. Government agencies—or, more precisely, people who work at the agencies and who have some control over what the agencies do—pursue goals of some sort. Whether it is the Pentagon or a state department of corrections, a government agency (or, more precisely, its high-level officials) does obtain some benefit from its service provision.

Moreover, agencies are not the only actors. The employees of the agencies, through their unions, also enjoy some benefit from public provision of the service, and they also participate in political advocacy. The challenge is to determine who the relevant actors are and what benefits they might plausibly seek to maximize. This is what I try to do, informally, in Parts II and III for corrections agencies and corrections officers unions.

The model implies, at a minimum, that some amount of privatization will decrease advocacy, for two reasons. The first reason is that, as long as the level of privatization does not exceed a certain critical threshold, the public sector will dominate the entire private sector (in terms of “real” share). Therefore, the model predicts that the whole private sector’s advocacy would be zero. The second reason is that as privatization increases, the size of the public sector falls, and thus the aggregate benefits of service provision to the public sector likewise fall. Because the public sector is smaller than it would be without privatization, its advocacy will fall accordingly.

How far can we continue to privatize before advocacy stops falling? As privatization increases, the second step always holds—by definition, privatization shrinks the size of the public sector. The first step, however, does not always hold for large enough levels of privatization. Obviously, at a certain point, the private sector can come to dominate the public sector. Then the private sector will do all of the advocacy, with the public sector acting as a free rider. From then on privatization would increase advocacy. The level of privatization at which advocacy stops falling is a threshold that we may call an “advocacy-minimizing privatization level.”

For instance—going back to the graphical model above—suppose our two firms benefit identically from having a given proportion of the industry. Then the advocacy-minimizing breakup is an equal split of the industry. This is because the amount of advocacy spending depends on the dominant firm’s real share, and the lowest possible real share of a dominant firm occurs when the dominant firm’s real share equals that of the smallest firm.

If a split in the industry creates a splinter firm that is twice as profitable as the incumbent firm, or perhaps twice as slick, then the advocacy-minimizing split is 67%–33%, again allocating each firm an equal real stake in the system. The splinter firm is half as large but twice as profitable, so again the dominant firm (in terms of “real” share) is as small as
it can possibly get. Conversely, if the splinter firm is only half as profitable as the incumbent firm, the advocacy-minimizing split is 33%-67%.

This concept becomes useful in the next Part. In Part II, I argue that the public sector share of total benefit is currently much larger than the private sector share—first, because the public sector still has a larger industry share, and second, because private firms are subject to a more competitive regime, so their profits are fairly low. Thus, the advocacy-minimizing level of privatization is probably quite high. Within this model, privatization would have to reach very high levels before pro-incarceration advocacy starts rising.

II. APPLYING THIS MODEL TO THE REAL WORLD

The model in Part I provides the intuition behind the story of industry-expanding lobbying after privatization. As privatization is introduced into an industry, the dominant sector—the public sector—advocates less than it used to, because it captures a smaller proportion of the benefits of its advocacy. The private sector is happy to free ride off the dominant sector’s contributions.

A. Different Kinds of Lobbying in the Real World

This basic result—that, by fragmenting an industry, one can reduce that industry’s political advocacy to increase its market—is also consistent with empirical studies on the relationship between industry concentration and lobbying.

In general, industry concentration can have two opposing effects on lobbying. On the one hand, a concentrated industry may be able to more easily overcome its collective action problems, so we might expect lobbying to increase as concentration increases. This factor is definitely relevant for industry-expanding lobbying. But consider another type of lobbying—anticompetitive lobbying, which seeks to regulate the market (for instance, through entry restrictions) to allow existing firms to charge above-market prices. Firms in a more concentrated industry can more easily suppress competition in the product market (either by just charging supracompetitive prices or by cooperating to change monopoly prices), so they have less need to do so through lobbying. They can raise prices above market levels all by themselves by directly using anticompetitive methods, so we might expect that a highly concentrated industry would have less to gain from anticompetitive lobbying than would a more competitive one.

Studies that do not disentangle these two effects can come up with results in either direction. One study found a positive effect of concentration on industry contributions, while another found that the percentage of firms with political action committees first rises and then falls as concentration increases.

This Article, though, focuses only on advocacy for reforms that increase the size of the industry, and not on advocacy for reforms that would squelch competition in the industry—since it is primarily the first sort of advocacy that privatization critics urge is illegitimate. So only the first of these forces comes into play here.
B. What Does the Model Predict About Prisons?

Now let us apply the theory to a real-world industry subject to the “political influence” critique of privatization: prisons. When I speak of “pro-incarceration advocacy,” I use the term “advocacy” broadly to include any use of political influence, licit or illicit, including endorsements, political contributions, lobbying, and bribes. And I use the term “incarceration” as shorthand to include the criminalization of a greater range of behavior, more active enforcement, greater reliance on imprisonment, longer sentences, and less parole—anything that ultimately increases person-years in prison. Thus, endorsing a politician for being “tough on crime,” donating money to a “Three Strikes” initiative, or testifying in favor of a “truth in sentencing” law all presumptively count as advocating incarceration.

Consider the main political actors in the prison industry: the private prison firms and the public corrections officers union. Without privatization, the public sector is the monopoly provider of prison services, and the corrections officers union enjoys the benefits that flow from serving the whole system. Now suppose that part of the system is privatized. At first, the public sector is clearly the dominant sector, that is, the sector with the largest proportion of total benefits from provision of the service. While the public sector’s proportion has gone down slightly from 100% of the industry, the private sector is still quite small. Because the public sector has shrunk, it is less willing than it used to be to spend money on reforms that would increase the size of the prison pie. Because the private sector is tiny, it acts as a free rider.

Privatization will always have this effect in the model presented above, provided the public sector stays the dominant sector. Any reform that shrinks the dominant sector will reduce industry-expanding advocacy. As it happens, this proviso is true in the case of prisons. At current levels of privatization, the public sector both has a larger industry share and extracts more benefit from the system than does the private sector.

We can easily perform some rough estimates to verify this:

- **Industry share.** The private sector has a smaller share of the industry. Of the 1.5 million prisoners under the jurisdiction of federal or state adult correctional authorities in 2004, 7% were held in private facilities. This includes 14% of federal prisoners and 6% of state prisoners. Among the thirty-four states with at least some privatization, the median percentage of private prisons was 8-9%. If we are interested in the private share of marginal prisoners—that is, how likely a prisoner is to go to a private prison if he is convicted today—the private share becomes larger, mainly because private firms have absorbed much of the recent growth in federal incarceration. A reasonable estimate of the private share of marginal prisoners over the period 2000-2005 yields 6% for state systems, 54% for the federal system, and 22% overall.

- **Private sector profitability.** The profits of the private sector are low. If the industry were perfectly competitive—like in textbook models of perfect competition—every firm would make zero economic profit. “Economic profits”
measures how profitable a company is relative to other ways of investing one’s money. Thus, “zero economic profits” does not mean that firms are not making money, but rather that all firms are doing as well as the rest of the market. In such a (hypothetical!) world, firms would not care whether their market were growing or shrinking, because they would be indifferent between running prisons and putting their money into the stock market. This is, of course, somewhat unrealistic: the prison industry is oligopolistic, not perfectly competitive, so prison firms do make some profit. But their profits are not high: 10% would be a generous estimate of prison firms’ profitability.

- **Public sector rents.** Public sector correctional officers, on the other hand, benefit substantially from public provision of prisons, because their wages are quite a bit above—about 30-65% higher than—what corrections officers make in the private sector. This is a lot of money, because wages are about 60-80% of most prisons’ operating expenses.

These numbers are meant to be merely suggestive, not rigorous. I make the assumptions—oversimplified but common in the economic literature on firms and unions—that firms maximize profits and that unions maximize total “union rents” (that is, here, the difference between public sector and private sector wages, times the size of the public sector). Trying to put these numbers together more or less rigorously requires a fair amount of algebra, which I provide elsewhere. But it should be intuitively plausible that our public-sector actors extract substantially more benefit from any given prison than do private firms. It is likewise clear that the public-sector unions have a greater share of the industry than do private firms.

Thus, overall, the public-sector actors enjoy a greater benefit from prison provision than the private-sector actors do, perhaps by an order of magnitude. This model predicts that the public-sector unions should be doing all of the pro-incarceration advocacy, and the private firms should be entirely free-riding.

C. Is This Realistic?

The theoretical model predicts that if an industry is divided, the actor that enjoys the greatest total benefit will foot the bill for all of the industry-expanding political advocacy, and the smaller sector will free ride. The rough estimates above suggest that, in the prison context, the sector with the greater benefit is the public sector, which not only still has a greater share of the industry but also benefits more from any given project. Therefore, we should expect to see pro-incarceration advocacy coming from the public, not the private, sector. Moreover, privatization reduces the public sector’s share of total benefits. So, at current levels, privatization should cause total pro-incarceration advocacy to decrease.

One may wonder about the realism of simple, highly stylized models. Will the private sector really do zero advocacy? Whatever the general merits of such skepticism, in this particular case the simple model may be close to true.
The next Subparts document what we know about prison industry advocacy. In brief, there is a lot of hard evidence of pro-incarceration advocacy by public corrections officers unions (though a small part of union advocacy also cuts the other way). On this issue, they are opposed to most departments of corrections, which advocate in favor of alternatives to incarceration. But there is virtually no evidence of private sector pro-incarceration advocacy. This may simply mean that the private sector advocates incarceration secretly. But, in light of the theory, it may be more plausible that the private sector simply is a free rider, saving its political advocacy for policy areas where the public good aspect is less severe—pro-privatization advocacy.

Even if one disagrees with the preceding sentence, this model need not be realistic in a literal sense. Advocacy need not be an entirely public good, and the smaller actors in the industry need not be complete free riders. The point is merely that these assumptions are plausible, perhaps even likely. Advocacy has some public-good aspects, and free riding happens to some extent in the world. If people act enough like this model, privatization will still, on balance, reduce total pro-incarceration advocacy.

This plausible scenario rebuts the simple anti-privatization claim that privatization does increase pro-incarceration advocacy. (The extended models presented later on, in which the effect of privatization on advocacy is ambiguous, further rebut the simple unidirectional claim.) This scenario also points out a potential irony in the position of some incarceration opponents who, so as to avoid “reinforc[ing] the incarceration boom by introducing the profit motive into incarceration,” would make common cause with public corrections officers unions, who concededly are active lobbyists for incarceration.

D. Public Corrections Officers Unions

In 1987, E.S. Savas, a supporter of privatization, dismissed the claim that private firms advocate incarceration by noting that “[i]f this argument was sound . . . prison officials, guards, and their unions presumably would act in the same manner for the same reasons. This, however, is not the case.”

Whether this was true even back then is questionable. At one time, corrections officials were politically aligned with liberal groups, but by the 1970s correctional unions were already advocating incarceration.

This activism continues today—for instance, through the California Correctional Peace Officers Association (CCPOA). The CCPOA gives twice as much in political contributions as the California Teachers Association, though it is only one-tenth the size—only the California Medical Association gives more in the state. CCPOA spends over $7.5 million per year on political activities. It contributes to political parties, political events, and debates; it gives money directly to candidates; it hires lobbyists, public relations firms, and polling groups.

Many of its contributions are impossible to trace back to any particular agenda item: since the union also opposes privatization, favors higher wages, and has positions on other issues, it is just as plausible that the contributions are for those other purposes.
But many of its contributions are directly pro-incarceration. It gave over $100,000 to California’s Three Strikes initiative, Proposition 184 in 1994, making it the second-largest contributor. It gave at least $75,000 to the opponents of Proposition 36, the 2000 initiative that replaced incarceration with substance abuse treatment for certain nonviolent offenders. From 1998 to 2000 it gave over $120,000 to crime victims’ groups, who present a more sympathetic face to the public in their pro-incarceration advocacy. It spent over $1 million to help defeat Proposition 66, the 2004 initiative that would have limited the crimes that triggered a life sentence under the Three Strikes law. And in 2005, it killed Governor Schwarzenegger’s plan to “reduce the prison population by as much as 20,000, mainly through a program that diverted parole violators into rehabilitation efforts: drug programs, halfway houses and home detention.” CCPOA does not always favor increasing incarceration, but the bulk of its advocacy has been in this direction.

That corrections officer unions benefit from increasing incarceration is plausible. Dan Pens has quoted CCPOA member Lt. Kevin Peters as saying:

> You can get a job anywhere. *This* is a career. And with the upward mobility and rapid expansion of the department, there are opportunities for the people who are [already] correction staff, and opportunities for the general public to become correctional officers. We’ve gone from 12 institutions to 28 in 12 years, and with “Three Strikes” and the overcrowding we’re going to experience with that, we’re going to need to build at least three prisons a year for the next five years. Each one of those institutions will take approximately 1,000 employees.

This is not just a story about California. Though corrections officers unions outside of California are nowhere near as active as the CCPOA, many of them do advocate incarceration. (As I note below, *everything* is bigger in California: while private prison firms make political contributions nationwide, they, too, spend more in California.) The correctional wing of Florida’s police and corrections officers union has endorsed candidates for being tough on crime. The Michigan corrections officers union has opposed boot camp proposals. The New York City corrections officers union endorsed Governor Pataki because he ended parole for violent felons. The New York State corrections officers union is said to have stymied efforts to overhaul mandatory minimum sentences. And the Rhode Island corrections officers union endorsed a candidate for his prosecutorial record and position in favor of tougher criminal penalties. (I am not considering the more usual demands for tougher penalties for criminals who commit crimes while in prison—a particularly salient issue for corrections officers, who are often victims of such crimes.)

Some corrections officers unions are combined with police unions, for instance in Florida and New Jersey. So except where (as in Florida) the corrections officers’ wing has been independently politically involved, these combined unions’ advocacy cannot be traced directly to corrections officers.

In some states, corrections officers are also affiliated with American Federation of State, County, and Municipal Employees (AFSCME), the general public employees union. AFSCME Corrections United represents 60,000 corrections officers and 23,000
corrections employees nationwide. It is plausible that corrections officers’ concerns would be swamped by the potentially contrary concerns of public employees as a whole (who tend to be fairly liberal). And, indeed, the evidence that AFSCME has advocated incarceration is weak. AFSCME has advocated alternatives to incarceration, and the national organization has advocated legalizing medical marijuana (though of course this would only account for a tiny proportion of crime). The Oklahoma public employees union—also a general union—has also advocated alternatives to incarceration.

E. Private Prison Firms

Private prison firms depend, for their livelihood, on two policies: privatization and incarceration. Indeed, they admit as much to the world, in their annual reports filed with the SEC. As to privatization, The GEO Group, the second largest private prison firm, explains that “[p]ublic resistance to privatization of correctional and detention facilities could result in our inability to obtain new contracts or the loss of existing contracts, which could have a material adverse effect on our business . . . .” As to incarceration, GEO candidly remarks:

[A]ny changes with respect to the decriminalization of drugs and controlled substances or a loosening of immigration laws could affect the number of persons arrested, convicted, sentenced and incarcerated, thereby potentially reducing demand for correctional facilities to house them. Similarly, reductions in crime rates could lead to reductions in arrests, convictions and sentences requiring incarceration at correctional facilities.

Similar statements are easily available in prison firms’ public filings. It is thus natural to suspect that prison firms may advocate both privatization and incarceration in the public square. Their political advocacy—which is extensive—mainly takes the forms of contributions to politicians and participation in the American Legislative Exchange Council (a conservative organization that drafts model legislation), though they also testify before Congress and present arguments in the popular press. But, while it is clear that these firms advocate privatization, it is unclear that they advocate incarceration to any significant extent.

Most of the evidence of advocacy specifically in favor of incarceration has been speculative. Some writers state that it does not happen or that “the impact of any private prisons lobby is, for the foreseeable future, likely to be peripheral at best,” while others who are concerned about the prospect describe the concern but stop short of claiming that it does or will happen.

I noted above that the general contributions of corrections officers unions cannot be traced back to any specific goal, like pro-incarceration advocacy. Some commentators note private prison firms’ advocacy without distinguishing between pro-privatization and pro-incarceration advocacy, but this blanket approach is a mistake, unless one is attacking all political involvement by prison firms. Generalized contributions to candidates, unlike targeted activities like contributions to single-issue voter initiative campaigns, are mute. The industry’s contributions to politicians may not be pro-incarceration at all; or they
may be multipurpose, for privatization as well as for incarceration. This is an important distinction, as merely advocating increased privatization arguably raises quite different concerns than advocating changes in the criminal law itself, and may not implicate the same sorts of "legitimacy" values.

Since the industry’s public statements virtually all relate to favoring privatization, there is little hard evidence on the basis of which to attribute part of their political contributions to a pro-incarceration motive. Indeed, the Association of Private Correctional and Treatment Organizations (APCTO), the industry’s trade group, speaking for its member firms, denies that the industry lobbies for increased penalties:

Individually and as an Association, we do not lobby in favor of longer sentences, so-called “three-strikes” laws, or other legislation which could result in an increase in the jail or prison population. To the contrary, the Association and its member companies encourage the use of appropriate alternatives to incarceration; provide inmates with treatment, education and rehabilitative services designed to positively impact and reduce recidivism rates; and encourage effective transitional programs for offenders upon release.

APCTO frequently endorses alternatives to incarceration, treatment programs, and other measures to reduce recidivism. Its executive director recently suggested in the *Denver Post* that to alleviate prison overcrowding, Colorado should “[l]ook to alternatives to incarceration that can provide treatment and rehabilitative programs to first-time, nonviolent drug and alcohol offenders,” “[r]educe recidivism by investing in the treatment, education and rehabilitation that offenders need to be successful when they leave prison,” and “[i]ncrease the likelihood that released inmates will not re-offend by providing substantive transitional programs to help released inmates adjust to the community outside the walls of prison.” (He made similar recommendations regarding Ohio in the *Cincinnati Post.* He also suggested in the *Fort Pierce Tribune* and the *Palm Beach Post* that Florida should invest more in juvenile justice services in order to reduce the adult prison population in the long run. (He noted that APCTO’s member companies mostly provide adult incarceration services, though some would like to expand their juvenile programs.)

Even if one ignores the industry association’s official statement as self-serving and dismisses their anti-incarceration positions as public relations, at most generalized political contributions are “soft evidence” of pro-incarceration advocacy. The most we can say empirically based on such evidence is that maybe pro-incarceration lobbying happens and maybe it does not. Perhaps the hard evidence is missing because the industry covers its tracks; or perhaps the hard evidence is missing because there is nothing to cover up.

Prison firms also participate in the American Legislative Exchange Council, an influential conservative organization that drafts model legislation. Both Corrections Corp. of America (CCA) and the former Wackenhut Corp. (now called the GEO Group) have been members of ALEC (and they and Sodexho Marriott, a major CCA stockholder, are prominent corporate funders of ALEC), and, over the years, at least CCA has
participated in (and two of its executives have chaired) ALEC’s Criminal Justice Task Force, which drafted, among other things, a “Truth in Sentencing Act” and a “Habitual Violent Offender Incarceration Act.”

The inner workings of ALEC are hazy, and indeed, some commentators argue that the private prison industry expressly seeks out channels that are “conveniently out of public view” and “behind closed doors” to promote its pro-incarceration agenda. The trouble with this view is that we can also presume that prison firms work within ALEC on privatization issues: Prison privatization is one of the “major issues” of the very same Criminal Justice Task Force; the Task Force has a Subcommittee on Private Prisons and has a model “Housing Out-of-State Prisoners in a Private Prison Act”; and CCA is known to have talked to the Task Force on the subject. Therefore, this, too, is “soft” evidence; we do not know that they also work on sentencing or incarceration issues. Indeed, CCA asserts that it has not participated in, voted on, or endorsed any stand on model legislation for sentencing or crime policies within ALEC.

Apparently, the only CCA official to have ever publicly taken a stand on sentencing is J. Michael Quinlan, formerly Director of the Federal Bureau of Prisons and now a CCA Senior Vice President, who, after he joined CCA in 1993, told a House subcommittee that mandatory minimum sentences “are unnecessary for non-violent, non-serious offenses” and “pose[] a severe threat to prison discipline and management.”

So far, I have found a single piece of evidence of arguably pro-incarceration advocacy by a private firm. In 1995, Wackenhut chairman Timothy P. Cole testified in favor of certain amendments to the Violent Crime Control and Law Enforcement Act of 1994. The main point of his testimony was to propose additional provisions (1) making clear that prison grants under the 1994 Act would “help pay for the entire range of correctional services states can provide in-house or under contract” (not merely for “alternative correctional facilities”), (2) requiring states to “show that they have all the necessary legislative authority to embark upon a comprehensive, integrated program and that they will employ the best technology at the lowest cost” (presumably to boost privatization), (3) directing the Attorney General to “give top priority to the construction of larger, ‘harder’ [i.e., higher-level security] facilities,” and (4) directing the Attorney General to give priority to states with “an executive body dedicated to the review and consideration of privatization.” During this testimony, he said the following:

- “Our proposed amendment . . . would help to assure that these grants will help the states incarcerate more violent criminals and not make the state governments more dependent on federal tax dollars in the long term.”

- “By passing ‘truth-in-sentencing’ laws, states have begun to restore a fundamental sense of justice and fairness to our system of crime and punishment.”

- “The new grant program [under the 1994 Act, without the proposed amendments] is available for ‘alternative correctional facilities’ and does not recognize the urgent need for more cells in secure facilities.”
• “Current law encourages billions to be spent on new or retrofitted facilities that are not large enough, secure enough or efficient enough to keep the maximum number of violent criminals in prison for the least cost.”

This is not great evidence—Cole was primarily advocating funding priorities and privatization-friendly decisionmaking. Cole’s request to divert money from alternative facilities, his kind words for truth-in-sentencing laws, and his positive attitude toward locking up violent criminals are hardly a pro-incarceration smoking gun. But this is the best I have found. Private prison firms may have made other statements and taken other public positions that are arguably pro-incarceration, but I have not found any, and to my knowledge, privatization critics have not brought them to light.

\[ F. \text{ Sometimes, No Smoke Means No Fire} \]

As noted above, there is little hard evidence that private firms advocate stricter criminal law at all. Perhaps they do so secretly, in which case this simple model may be entirely unrealistic. Or perhaps this simple model is basically right, and the private firms are actually spending their money on a form of advocacy where the public good aspect is not important—\textit{pro-privatization advocacy}.

Pro-privatization advocacy is an area where, obviously, the private sector cannot free ride off the public sector, since the public sector is their enemy on that issue. If the private firms cooperate with each other, they reap all the benefits of their pro-privatization advocacy. Even if they do not cooperate with each other, an individual firm’s pro-privatization contribution may benefit it directly to the extent that the contribution (perhaps improperly) increases the likelihood that the firm will obtain a particular contract.

In real life, of course, money may be multi-purpose. So far, I have treated “mute” campaign expenditures as though they were for some purpose—either privatization or incarceration—that was known to the donor but unknown to us. In fact, they could be for “access” to the candidate, which can be used at any time after the candidate prevails. The model is general enough to accommodate this framework. At some point, donors will try to call in a favor. Favors cost something in terms of “political capital,” and political capital is scarce: calling in one favor makes it harder to call in another favor. At the point where donors have to determine what to ask for, we are back in the previous model.

The “access” framework has thus only postponed the applicability of the model until after the election. One would still predict, under this model, that the smaller donors would prefer to spend their capital supporting something with more of a private-good component, like privatization, and leave the pro-incarceration advocacy to the dominant actor. And this may in fact be what happens.

\[ \text{III. OF FIRMS, UNIONS, AND COOPERATION} \]

This Part elaborates on two curlicues of the theory. Subpart A explains why I have focused only on public-sector unions and private firms. In short, I have done so because
the other potential prison-based actors do not participate in pro-incarceration advocacy. Private-sector workers are not unionized, which makes it hard for them to act collectively; and public departments of corrections actually want fewer prisoners. Subpart B explains why I have assumed that the private firms act as a bloc instead of competing with each other, or, at the opposite extreme, cooperating with the public sector in a grand prison coalition. In short, I have made this assumption because cooperation within a fairly concentrated oligopoly is not that difficult: firms interact with each other a lot and have ample opportunity to punish each other for non-cooperative behavior. Moreover, private-sector firms interact with each other more than they do with the public sector, so enforcing cooperation across the whole prison industry would be more difficult than merely doing so among private firms. However, it turns out that how the industry cooperates, or whether it cooperates at all, does not make much of a difference for the main result.

A. Why Focus on Public-Sector Unions and Private Firms?

One might ask, at this point, why I have focused primarily on two apparently asymmetrical groups: the private sector firms and the public sector employees. What about the other two obvious candidates—the employees of those firms, or the employers of those public guards?

In principle, it is unclear a priori who would want to lobby, and so a case-by-case analysis of the incentives of the various parties’ incentives is necessary. In this case, my choice of actors was inspired by the state of the evidence and the debate: public corrections officers unions, especially in California, are known to engage in pro-incarceration advocacy; and private prison firms are alleged to do so. But let us also think about this theoretically.

First, let us consider the workers. No single employee has enough of a stake in the system to benefit from spending resources on advocacy to help his industry. We should only expect workers to be a significant political force if they can enforce some sort of collective action by punishing their own free riders. The easiest way to accomplish this is to require membership in or contribution to a union, which then lobbies out of members’ dues. Private corrections officers are not unionized in most states, which explains why they have not been observed lobbying.

As I explain above, I assume that unions represent their members and seek to maximize total union rents—the difference between union and non-union wages, times the size of the public sector. The prediction that such unions would seek to increase the size of their sector is straightforward (though not automatic). A larger sector may mean a more powerful union and therefore potentially higher wages, benefits, or job security down the road (and perhaps—to introduce agency costs for a moment—perks for union officials). It is possible that unions may sometimes oppose expansion of their industry—for instance, if increases in prisoners made corrections officers worse off because they were not accompanied by compensating wage or staff increases. Anti-expansionary lobbying may occur in some industries, but it apparently does not occur in the prison industry. The public corrections officers unions seem, so far, strong enough that they believe that an
increased flow of prisoners will not make them worse off (even if budgets are tight elsewhere in the system).

Now, let us consider the employers. Some private prison firms also run alternatives to incarceration, so it is not obvious that they would advocate an increased emphasis on imprisonment. Still, they may benefit from the other elements I have included in the term “incarceration”: increased illegalization, increased law enforcement, and longer sentences (once the imprisonment decision has already been made). Though increased incarceration may also increase costs for private firms, they have a built-in protection against too much deterioration in their position: they do not have to bid on a contract unless they anticipate making enough profit. So it is not implausible that private firms would benefit from incarceration; though of course (as explained in the previous Parts) their willingness to spend money to increase incarceration depends in part on how profitable they are.

What about the public employers, the departments of corrections? They are not players in the pro-incarceration advocacy game for a simple reason: generally, they favor alternatives to incarceration.

The Alabama Department of Corrections (DOC) commissioner has advocated sentencing reform, community correction programs, and other measures to “reverse the prison population growth trend.” The head of the Illinois DOC advocates reentry programs that would lower the prison population by countering the “awful, vicious cycle” by which recidivist parolees are re-incarcerated “before the ink is dry on their parole papers.” The Michigan DOC director concerns herself with measures to reduce the prison population and thus delay the day the state runs out of funded capacity for prison beds. The Montana DOC director candidly tells crowds that “[p]rison isn’t working,” and his department considers measures to reduce the prison population and increase community corrections. The New Mexico Corrections Department is focusing on using early parole to control its prison population. The North Carolina Division of Community Corrections advocates redirecting non-trafficking drug users from prison to “intermediate programs.” Ohio lawmakers complain about the high costs of mandatory minimum sentences. The Pennsylvania DOC is implementing programs “aimed at diverting less serious offenders from prison” to “free-up prison space needed for more serious offenders.” The Washington DOC secretary “is a big believer in work-release programs.” And the Wisconsin DOC secretary advocates focusing on “prevention and treatment in addition to effective law enforcement.”

This makes some sense: while it is commonly thought that agencies want to aggrandize themselves, that intuition is only a special case of a more general belief that agency officials act in their own self-interest and that their self-interest tends to be aligned with the size and power of their agencies. Increasing prisoners without a corresponding budget increase to match the increasing cost of incarceration (a cost that includes corrections officers’ salaries, as well as health care and other factors) can easily make prison officials worse off. Thus, the interests of departments of corrections may not be aligned with those of corrections officers and their unions. Moreover, DOCs run both prisons and many alternative programs, so even if more inmates mean more power for the DOCs, it makes sense that the DOCs would want to handle those inmates in cheaper ways than
incarceration. Thus, it is not surprising to find prison systems arguing for alternatives to incarceration in a time of tight budgets.

**B. Who Cooperates with Whom?**

The discussion in the previous Parts of how the 10% firm acts and the profits of “the industry” assumes that the private sector, in deciding how much to spend, acts as a bloc: the private firms all cooperate with each other, but do not cooperate with the public sector. This is possible, but it is not the only imaginable story. I could have made either of two other, more extreme assumptions. First, there could be no cooperation at all—all the firms could be acting independently and competing with each other. Second, there could be total cooperation—all the firms could be cooperating not only with each other but also with the public sector. This section explores the implications of these alternative assumptions and tentatively defends my decision to adopt the intermediate assumption of cooperation within the private sector but not with the public sector. In the end, however, making different assumptions would not significantly alter the conclusions.

If all firms act independently, the relevant shares are even less than indicated above. In 1999, CCA had a bit over half the market, Wackenhut (now the GEO Group) had about a quarter, Management & Training Corp. had about 5-8%, Cornell Corrections, Inc. and Correctional Services Corp. each had about 5-6%, CiviGenics, Inc. had about 2-3%, and a handful of other firms had under 1%.

So, while the private-sector share in some state may be 10%, that number is irrelevant if all firms act independently. The relevant shares may be, for instance, 6% for CCA, 2% for The GEO Group, 1% for Management & Training Corp., and 1% for Cornell Corrections, leaving 90% for the public sector. The assumption of independent firms makes public sector domination even more probable.

Now consider the opposite assumption—that everyone cooperates. A single prison industry bloc would choose an optimal total amount to maximize total industry benefit. Because the actors are still formally separate, they would also choose some way to allocate the contributions among themselves.

If private firms had the same benefit per prison as the public-sector union, then total cooperation would be indistinguishable from monopoly: total industry benefit would be the same before and after privatization, thus the strategy that chooses contribution amounts to maximize that benefit would likewise be the same.

However, as I argue above, private firms are not terribly profitable, while public-sector unions have significant public-sector wage premiums to protect. By replacing part of the public sector with a relatively unprofitable private sector, privatization actually decreases the industry’s total benefit. Therefore, even under total cooperation, there is less to maximize. Expenditures on pro-incarceration advocacy are thus less productive (just as if there were a tax rate on industry revenues), and expenditures on advocacy still go down under privatization.
How can we tell which form of cooperation is most likely? Not being able to find explicit cooperation does not mean anything: the cooperation may just be tacit. Observing the private industry’s trade association, APCTO, also fails to answer the question, because although trade groups may provide a forum for discussing common lobbying strategies, talk is cheap and many trade groups are ineffective. In particular, APCTO does not seem to fulfill much of a coordinating function, as firms do their own lobbying and most of their own advocacy.

Even observing some actual lobbying by the major firms does not answer the question. As I noted above, they may all be lobbying for privatization, which has a strong private-good component, since a firm’s contributions may increase the probability that it gets a project in the future.

On theoretical grounds, it seems at least plausible that the private firms would cooperate among themselves. They are repeat players in a long-term process, which includes both political advocacy and bidding on actual prison projects. Therefore, there is ample opportunity for private firms to enforce a regime of cooperative behavior. If firms free ride off each other in their advocacy expenditures, their fellows could punish them in the future in any number of ways—for instance, by never cooperating on campaign spending anymore, by bidding aggressively in prison auctions, or by bidding aggressively only in certain markets.

By contrast, public corrections officers unions may have fewer ways of punishing private firms. They do not bid against each other in the underlying auctions, so they cannot threaten to end cooperative behavior. Public unions are bitter political adversaries in the privatization advocacy world, so again there seems to be no preexisting cooperation that can be terminated. They can threaten to not cooperate anymore in pro-incarceration advocacy or to step up their anti-privatization advocacy, but this may not be as effective a threat.

For these reasons, I believe that cooperation among private prison firms is more likely than either, on the one hand, totally non-cooperative behavior or, on the other hand, totally cooperative behavior between the public and private sectors. However, because the ultimate results under any of the assumptions do not differ that much, which assumption we choose is not terribly important.

IV. COMPLICATING THE MODEL

The theoretical model in Part I was highly simplified. It is, therefore, not surprising that its central prediction—that smaller actors would do no advocacy at all, and that privatization (up to the “advocacy-minimizing” level) would unambiguously decrease the level of industry-expanding advocacy—was also simple. This Part complicates the model in various ways.

In Subpart A, I drop the assumption that money only buys victory for a given reform or candidate and introduce the possibility that money can also change the substance of the reform or the candidate’s position. This does not significantly alter the conclusion. In
Subpart B, I drop the assumption that anti-incarceration political advocacy is fixed. I find that pro-incarceration advocacy still falls with privatization, though the effect of privatization on anti-incarceration advocacy is ambiguous.

The following two Subparts show how privatization may have an ambiguous effect on pro-incarceration advocacy. In Subpart C, I relax the assumption that all money is fungible and that only the total amount of money in the pot matters. Once we allow public-sector money and private-sector money to have different effects, privatization has an ambiguous effect on total pro-incarceration advocacy: private advocacy rises, but public advocacy falls. In Subpart D, I introduce the possibility that the pattern of privatization, as we observe it today, is already the result of a political process where strong unions have successfully opposed privatization while weak unions have not. I find that exogenously increasing privatization in such an environment would likewise have an ambiguous effect on pro-incarceration advocacy; the effect depends on the correlation between actors’ influence in privatization politics and their influence in incarceration politics.

The bottom line is that, if one wants to argue that privatization will increase pro-incarceration advocacy, one must argue either, from outside the model, that the model is wrong, or, from inside the model, why privatization would increase private-sector advocacy more than it would decrease public-sector advocacy.

A. Allowing Money to Change Candidates’ Positions

So far, I have taken the political agenda as given: I did not explain the source of the proposed reform. Thus, I have assumed that money is important because it buys victory—for instance, by persuading voters of the benefits of the policy or the merit of the candidate. But money can also affect the agenda by changing candidates’ positions, inducing the sponsors of voter initiatives to propose a different initiative, and so on.

When money affects candidates’ agendas (but the other assumptions are unchanged), the analysis is similar. Suppose you are considering whether to contribute to place an initiative on the ballot. The initiative is supported by some group or other, but for specificity, assume a politician sponsors it. This politician may be fairly pro-incarceration himself, but he is limited in how strict an initiative he can propose: his effort will fail unless the median voter, whose views control the outcome of the election, prefers the proposal over the status quo. However, before the substance of the initiative is set in stone, you can move him in a more pro-incarceration direction if—by offering him money to pay for persuasive advertising—you offer him the possibility to also move the median voter.

A monetary contribution has the following effects:

1. Electoral influence.
   a. As before, you benefit because your contribution pays for persuasion, directly increasing the probability that the initiative prevails.

757
Draft — do not distribute!
b. But the contribution also moves the initiative in a more pro-incarceration direction, which cuts against the effect above.

2. **Substantive influence.** Finally, you benefit if the initiative prevails, because the policy is better for you than it would have been if you had not contributed.

It turns out that this complication to the model does not much change the underlying result. As a prison provider thinking about how much to contribute, you follow the same framework as before: you contribute until the benefit of an extra dollar is worth $1 to you. The benefit of an extra dollar is more complicated than it was in the earlier model because, in addition to encompassing the positive electoral influence effect, it now also includes the negative electoral influence effect, as well as the benefit of substantive influence. The basic idea, however, remains the same.

Now suppose, again, that the industry is split up into a 90% sector (you) and a 10% sector (them). As before, your benefits shrink to 90% of their previous level, so you now want to contribute until the benefit of an extra dollar to the industry is worth $1.11. As before, you contribute less than before the split, because having only 90% of the industry is like facing a 10% tax on benefits. Also as before, your competitor free rides off you, because when he takes your contribution level into account, an extra dollar in the pot is no longer worthwhile to him.

**B. Anti-Incarceration Advocacy**

This model focused only on pro-incarceration advocacy, taking the anti-incarceration advocacy as given. But clearly anti-incarceration advocacy exists, and it is plausible that the pro- and anti-incarceration forces respond strategically to each other’s expenditures. This suggests two questions.

First, one might wonder whether the existence of anti-incarceration advocacy changes my conclusions about the effect of privatization on pro-incarceration advocacy. It turns out that it does not: just as in the simple case, privatization makes pro-incarceration advocacy decrease, even when we consider interactions with anti-incarceration advocacy.

Second, one might wonder how privatization changes anti-incarceration advocacy. After all, some anti-incarceration advocacy is as plausibly self-interested as the prison providers’ pro-incarceration advocacy. For instance, Proposition 66, which would have limited California’s Three Strikes Law, was partly funded by “Sacramento businessman Jerry Keenan whose son Richard is serving time for manslaughter after crashing his car while driving drunk and killing two passengers.” Proposition 36, the drug treatment diversion initiative, was supported by dozens of drug treatment providers and seventeen medical and public health organizations, including the California Association of Alcoholism and Drug Abuse Counselors and the County Alcohol and Drug Program Administrators Association of California. And, as shown above, state DOCs generally advocate against incarceration. Perhaps those who are concerned about self-interest coloring people’s positions on criminal justice should be concerned about this self-interested anti-incarceration advocacy as well.
It turns out that the privatization-induced decrease in pro-incarceration advocacy has an indirect effect on anti-incarceration advocacy. Unfortunately, we cannot say anything a priori about the direction of this effect. On the one hand, pro-incarceration advocacy decreases the effectiveness of anti-incarceration advocacy by counteracting it. A decrease in pro-incarceration advocacy, therefore, makes anti-incarceration advocacy more effective, which would tend to increase it. On the other hand, a decrease in pro-incarceration advocacy also makes anti-incarceration advocacy less necessary, which would tend to decrease it. There is no theoretical way to know how these conflicting effects would balance out; but in principle, future research could answer the question empirically.

What this means normatively depends on one’s attitude toward anti-incarceration advocacy. If one opposes pro-incarceration advocacy because there is already too much incarceration, then there is nothing wrong, and perhaps everything right, with advocacy the other way.

On the other hand, if one opposes pro-incarceration advocacy because it is assumed to be self-interested, then perhaps anti-incarceration advocacy is just as bad if it comes from boot camps, halfway houses, drug treatment providers, and other presumptively self-interested parties. This model, which says nothing specific about total advocacy (either its amount or its effect), is thus normatively ambiguous.

C. Relaxing the Assumption of Fungible Money

Recall the main model presented in Part I, in particular Figures 4 and 5. A monopoly provider would have spent $1 million on advocacy, but under a 90-10 split, the 90% provider is unwilling to spend beyond the 900,000th dollar and the 10% provider is unwilling to spend beyond the 300,000th dollar; and so total advocacy falls to $900,000, with the dominant provider spending everything and the other one spending nothing.

The result that the smaller-share-of-total-benefit sector totally free rides off the efforts of the dominant sector was driven by the assumption that the probability of getting the change in policy only depended on the total amount of money in the pot. All advocacy was fungible. A dollar from a public actor had the same effect as a dollar from a private firm. This is not an implausible assumption. For instance, dollars are fungible in buying advertising, which increases the probability of a change. A politician may adopt the view of whatever “policy position” contributed the most to his war chest.

On the other hand, some alternate assumptions may also be plausible. For example, one group might be attractive only to Democrats, while another might be attractive only to Republicans. More generally, perhaps politicians are just sensitive to the variety of voices in a coalition, feeling (rightly or wrongly) that having a wide variety of groups shows that a policy has wide support. Then neither group’s contributions totally “crowd out” the other’s. Your 500,001st dollar still has less benefit than your 500,000th dollar—there are still decreasing marginal returns—but (unlike in the previous model) it does not have the same benefit as your first dollar added on to your competitor’s 500,000th. Therefore, the total free-riding effect from the simple model above no longer occurs. There are many
ways that private and public spending could interact. For instance, the effect of a public dollar could be the same regardless of the level of private spending, and vice versa. Or, alternatively, public and private spending could be complementary if politicians are eager to endorse a policy supported by actors from both the public and private sectors. This is an empirical question to be answered by future research.

In this context, privatization may have two opposing effects. First, it increases the private-sector share, so private-sector advocacy goes up. Second, it decreases the share of the public sector, so public-sector advocacy goes down. We cannot say anything a priori about whether the first effect outweighs the second. If we know some facts about public or private sector advocacy—for instance, if one sector is completely unpersuasive, while the other sector is slick and sympathetic—then we can hazard some predictions, but we cannot say anything without such empirical facts. Because the empirical effect of privatization is ambiguous, the normative effect of privatization is also ambiguous if one opposes pro-incarceration advocacy. Unless we can be specific about how different groups’ advocacy has different effects and how effective the groups are, it is impossible to say whether prison privatization increases or decreases self-interested pro-incarceration advocacy.

D. Strong and Weak Unions

Let us return to the point I made above that an industry’s effectiveness at advocacy is relevant to its “real” share for purposes of this analysis. For instance, if your competitor, with a 10% share, is twice as slick a lobbyist as you, meaning that his marginal dollars produce twice the benefit of yours, he will act as though his share is 20%.

Which way this cuts is not clear, as we do not know which sector is more effective at lobbying in favor of incarceration. The CCPOA, as we have seen, is highly effective, but corrections officers unions are much less active outside of California, and perhaps this is because they are less effective. It is hard to say how effective private prison firms are at lobbying in favor of incarceration, since, as we have seen, there is little evidence that they do this at all, and if they do it secretly, it is likewise hard to gauge how effective they are.

But let us suppose that one’s effectiveness at lobbying for incarceration is correlated with one’s effectiveness at lobbying for (or against) privatization. For simplicity’s sake, let us suppose that they are perfectly correlated. Consider the states with high levels of privatization. We may conclude that those states have high privatization because their corrections officers unions were not effective at opposing privatization; the private industry was just too strong for them. When that relatively “weak” public sector was partly displaced by a relatively “strong” private sector, a weak pro-incarceration voice was similarly displaced by a strong pro-incarceration voice. Pro-incarceration advocacy, then, may plausibly have increased.

Similarly, consider the states with low levels of prison privatization, like California (1.8% private in 2004), or no privatization at all, like New York or Rhode Island. The unions in those states, on this view, must have been stronger than the industry, or else we would see privatization there now. If privatization were introduced, total pro-
incarceration advocacy would go down; but privatization is unlikely to be introduced there, so we will not see that happen.

This is a story where—contrary to my implicit assumption so far—privatization is endogenous: the states where privatization has gained a foothold are not randomly chosen; rather, privatization emerges where corrections officers unions are weak and fails to emerge where the unions are strong. Thus, past privatization may have, on balance, increased pro-incarceration advocacy. If one could somehow eliminate prison privatization (despite the confluence of powerful political forces that established it to begin with), one would reestablish the rule of the ineffective corrections officers unions in those states where they were ineffective—to the benefit of those who oppose pro-incarceration advocacy. By a similar logic, one should introduce privatization where it is currently absent: if it is currently absent, it is because it was not a powerful enough political force to win on its own, which means it will also be an ineffective political force in fighting for incarceration.

In fact, the assumption here—that the effectiveness of pro-incarceration advocacy is perfectly correlated with the effectiveness of pro- or anti-privatization advocacy—implies that pro-incarceration advocacy is already as high as it can get, because the slick advocates, who were already slick enough to establish themselves in the industry, are now plying their slickness in the incarceration policy field. Adding a thumb to the privatization scales in either direction would tend to support the victory of the less persuasive party and would therefore reduce the total amount of pro-incarceration advocacy.

This story may be plausible, but it requires more fleshing out. For one thing, the assumption may not be right. Low-privatization states need not be high-union-strength states. While antipathy to privatization and the strength of public-sector unions are probably correlated, a very Democratic state may plausibly oppose privatization even if, for whatever reason, its corrections officers union is weak.

Moreover, actors in the prison industry may not be similarly effective in the privatization debate as in the incarceration debate. While one’s effectiveness at advocacy probably depends on one’s general characteristics, like goodwill, persuasiveness, and slickness, the specific subject matter of the advocacy also plays a big role. The incarceration debate is peopled by different interest groups than the privatization debate. For instance, prosecutors, police officers, victims’ rights groups, and rural communities are interested in incarceration policy but not so much in privatization policy. Conversely, prison privatization is a matter of interest even to interest groups without a direct interest in prisons, like, on one side, generalized public employee unions, and, on the other side, small-government advocates, who assume (probably sensibly enough) that a victory for privatization in any field is a victory for the general privatization movement. Moreover, the appeal of incarceration arguments, which connect to fears of drugs and crime and concerns over civil liberties, seems to have a very different source than the appeal of privatization arguments, which relate to taxes, spending, and the effectiveness of government services.
We are back, then, to a general state-by-state analysis. In the first set of models—where the effectiveness of advocacy only depended on the total amount of money in the pot—everything was driven by the dominant actor, where the term “dominant” also takes effectiveness into account. I have given arguments above as to why the private sector is currently probably the smaller actor. The “slickness adjustment” described here might change that in some places, but it is an empirical question. As is by now familiar, privatization still increases the private-sector share but decreases the public-sector share. This “slickness adjustment” may change the de facto shares of the different sectors, but it does not change the qualitative result. The effect of privatization is theoretically ambiguous.

CONCLUSION

I have explored how privatization affects the amount, or effectiveness, of economically self-interested pro-incarceration advocacy. For purposes of this Article, I have so far assumed, with the critics of prison privatization, that such advocacy is undesirable. But this assumption is highly questionable.

For one thing, members of an industry, whether public or private, who advocate a policy that benefits them are not necessarily motivated by self-interest, even unconsciously. When Don Novey, the president of CCPOA, says he just wants to lock up scumbags, perhaps we should take him at his word. The same goes when a DOJ official speaks of the need to fight “the scourge of child pornography,” when CACI says terrorism is “heinous,” when a leading environmental citizen-suit litigator argues against weakening the environmental laws whose monetary penalties fund its operations, or when doctors who perform abortions oppose abortion restrictions.

People who advocate a policy that benefits them or their industry may be acting out of naked self-interest; they may be deluded into believing their particular interest is the general interest; their participation in an industry may lead them to rightly appreciate their industry’s contribution to the public interest; they may have joined the industry because they were sympathetic to its interests; or maybe they just coincidentally believe that the policy is right.

Nor is even nakedly self-interested advocacy an obvious evil, even when prison policy is at stake. Some argue that optimal criminal law should reflect all interests, including the benefit to the criminal of committing the crime; and if this is right, prison providers’ self-interest is also relevant. Some see lobbying as a means by which groups provide their views to decisionmakers and the public and thus enrich democratic debate. Others may find it illegitimate, on democratic grounds, to even consider the substance of people’s future political advocacy in deciding whether to privatize.

And if, as still others believe, criminal policy should be judged by a substantive external standard—for instance, whether sentences are too long in an objective sense—one cannot specifically object to the effect of pro-incarceration advocacy on criminal law without first establishing that the effect would be substantively undesirable.
Nonetheless, if one believes that the effect of privatization on pro-incarceration advocacy is relevant, this Article has pointed out the inadequacies in the current formulation of the political influence challenge to privatization. My opinion, based on the above theory and evidence, is that privatization may not worsen any political influence problem, and might even alleviate it. The public goods model seems to describe many situations of political advocacy fairly well. The assumption of the principal model—that the probability of getting a policy change only depends on the total amount spent—likewise seems to be a good approximation for many situations, like initiative or election campaigns.

There is always room for more realistic theories. For instance, my analysis of what motivated public-sector unions, while based on assumptions common in the labor economics literature, was highly simplified. In assuming that private prison firms were profit-maximizing, I suppressed any analysis of agency costs within the firm. And my back-of-the-envelope estimate of the benefit of incarceration to the different sectors was just that—an estimate. Nor have I entertained the possibility that, when privatization is on the agenda, prison system actors spend more resources fighting over that, which might crowd out pro-incarceration advocacy. So my specific conclusions here are tentative. This Article is meant to stimulate and discipline further debate, not end it.

But what is not tentative is that this sort of analysis is necessary if one is to make the political influence argument properly, whether in the prison context or more generally. General assumptions will not do. As Mancur Olson (somewhat hyperbolically) observed, “the customary view that groups of individuals with common interests tend to further those common interests appears to have little if any merit.” Critics of privatization who have charged that privatization has increased (or will increase, or runs a substantial risk of increasing) industry-expanding advocacy have not explained what it is about the lobbying world that would make this happen. Either they are unambiguously wrong, or they are only right under a particular set of empirical assumptions that they must spell out.

One further note: If one opposes self-interested pro-incarceration advocacy, one may object at this point that this economic analysis does not exonerate private prisons. Rather, perhaps I have only shown that the entire system is corrupt, and perhaps I have unwittingly demonstrated that the only way out of this mess is to reject the “interest group model of politics” entirely as it applies to criminal justice policy.

Fair enough. If self-interested pro-incarceration lobbying is indeed undesirable, then perhaps the system is corrupt. But how does this translate into an argument against prison privatization? It is not enough to show that private prisons are part of the problem: removing one problem is not guaranteed to make things better when there are other problems around. As the models above have suggested, even if all this political advocacy is illegitimate, the existence of the private sector can reduce public-sector advocacy and may reduce total advocacy; eliminating the private sector may thus exacerbate the problem.

Nor is it only economists who oppose making the best the enemy of the good: as Rawls (no economist he) teaches, the analyst who makes specific policy recommendations in
our fallen world—not in the idealized world of “strict compliance” with the principles of justice that characterizes a “well-ordered society”—is acting in the realm of “nonideal theory,” which asks how the “long-term goal” dictated by ideal theory “might be achieved, or worked toward, usually in gradual steps. It looks for policies and courses of action that are morally permissible and politically possible as well as likely to be effective.”

Because nonideal theory requires that we ask about the real-world effectiveness of any reform, merely observing undesirable lobbying by the private sector will not support an argument against prison privatization unless, say, privatization actually (not speculatively) increases “the danger of . . . corrupting influence” or “compromise[s] further the possibility of legitimate punishment.”

If it turns out that privatization actually reduces pro-incarceration lobbying—if, with privatization, prisoners’ sentences are less influenced by improper factors than they otherwise would be—it is unclear that there is any “tension between the state’s use of private prisons and the demands of” liberal legitimacy. If “private prisons are by no means unique,” and if any prison provider, public or private, can lobby for incarceration, any “tension” has nothing to do with private prisons and everything to do with the crooked timber of humanity.

* * *

The same sort of analysis that I have conducted here on the prison industry can also be used to evaluate the claim that, say, buying weapons from defense contractors (rather than having the military make them in-house) will exacerbate pro-war lobbying. Since governmental providers of defense services—i.e., the military leadership—have, on some accounts, been notorious pro-war lobbyists throughout history, such a claim is not credible unless one can tell a plausible story about why any defense contractor lobbying will not crowd out some lobbying by the military itself; and doing this requires taking a position on what motivates the people at the Pentagon. The same goes for private attorneys general, private redevelopment corporations, private landfill operators, and the like. The result will not always be the same, and the political influence argument may turn out to be correct in some of these cases and incorrect in others. But this should be the structure of the argument.

The surprising moral of this story should not be that surprising. Indeed, the central insight here was also an important argument in favor of the antitrust laws. Discussing the conditions that preceded the enactment of those laws, William Howard Taft wrote that “business methods and plans . . . directed to . . . suppressing competition . . . had resulted in the building of great and powerful corporations which had, many of them, intervened in politics and through use of corrupt machines and bosses threatened us with a plutocracy.” The argument is plausible, and it is likewise plausible that privatization, by fragmenting an industry into at least two chunks (and more if private firms do not cooperate on advocacy), may similarly reduce that industry’s political power.
In a roundabout way, then, privatization is a form of antitrust, and antitrust is a form of campaign finance regulation. It may not be worthwhile to privatize industries—or break up large corporations—merely to reduce their political advocacy, but at the very least this may count as an unintended—and possibly happy—side effect of privatization that, if real, should be taken into account in future analysis.

§ 9.2.4. John J. DiIulio, Jr., *What’s Wrong with Private Prisons*, PUB. INTEREST, Summer 1988, at 66

**THE MORALITY OF PRIVATE PRISONS**

. . . Discussions of the morality of private corrections are almost as impoverished as the empirical literature on the subject. Two mistakes are often made. First, there is the substitution of empirical for normative criteria, giving rise to the notion that if it could be shown beyond a reasonable doubt that private involvement improves (or worsens) correctional services, lowers (or increases) correctional costs, and exceeds (or falls below) constitutional requirements, then the normative questions would thereby be resolved. . . .

The second mistake in discussions of the morality of private corrections is made less often by policymakers and practitioners than by academics who focus on the profit motive of the privatizers. They assume that the central moral problem of private corrections is not whether the authority to administer justice ought to be delegated by contract to private, nongovernmental parties, but whether the private contractors ought to be paid or to profit financially for their services. In effect, they worry less about whether the government’s responsibility to govern ends as the prison gates than whether (and how much) the gatekeepers get paid.

The question of whether it is ever right to profit from the misfortunes of criminals and their victims is a serious one. But it must be understood that there have been, and continue to be, several nonprofit, nonpublic corrections enterprises (for example, Florida’s Okeechobee School for Boys, run by the Eckerd Foundation). A properly constructed moral case for (or against) private involvement cannot turn simply, or even mainly, on whether the private contractor reaps huge or moderate profits—or none at all. By analogy, if one reasons that it is morally wrong (or right) for married men to have sex with women other than their wives, then it matters little (if at all) whether the act is done with high-priced call girls, cheap prostitutes, or the girl next door. . . .

**PRIVATE PRISONS TODAY**

. . . History . . . teaches us to be extremely wary of the oft-made claim that [private prisons will continue to have a good record] as long as contracts are detailed in accordance with the highest contemporary correctional standards. Writing and enforcing highly detailed contracts may help to guarantee accountability, but it does so only at the expense of administrative flexibility. A major selling point of the privatizers is that they
can be made innovative (and ultimately more efficient) than public providers of
correctional services. A contract that “counts nails,” however, makes it difficult for the
contractor to allocate resources freely, to make staffing changes, and so forth. As a public
corrections official with years of experience in these matters stated, “Either the
contractors will be allowed to run wild as they did in the old days, or we’ll make the
specifications, regulations, and monitoring so rigid that the firms will become as
bureaucratized and inefficient as we are—killing the goose before she lays any eggs.” . . .

One unchanged historical reality about American prisons and jails is that those who are
incarcerated in them come overwhelmingly from the lowest rungs of the socioeconomic
ladder. Criminologists and others may debate both the empirical and the moral issues
related to this fact, but none deny the simple truth that America’s prisons house
disproportionate numbers of poor, dark-skinned individuals. Historically, race and class
prejudices have combined with the low visibility of penal facilities to make the average
American unconcerned about what happens inside prisons and jails. It seems likely that
societal pressures against inmate abuse and political corruption will be at low ebb when
these largely underclass and minority populations of offenders are placed in nonpublic
hands. Opponents of privatization worry that the sigh of relief that public officials may
breathe after turning over all or part of their penal complex to private firms may presage a
“death rattle” for inmates’ legal and constitutional rights.

But such possibilities become less frightening in light of the single most important post-
1970 change in the operating environment of prisons and jails—namely, the rise of
judicial intervention into penal affairs. Since 1970, dozens of correctional systems have
fallen under court orders. And while there continues to be disagreement about the causes
and consequences of judicial intervention in this area, most experts agree that activist
state and federal judges have become a permanent part of the correctional landscape.
Although I harbor some serious reservations about the judges’ involvement, it is
impossible to deny that the courts can and do provide a salutary barrier against any repeat
of the kinds of unmitigated horrors that characterized the previous era of for-profit
corrections. If anything, the courts may prove especially eager to hold the feet of the
corrections corporations to their well-stoked judicial fires. . . .

ADMINISTRATION: PUBLIC VS. PRIVATE

Most claims that private corrections firms can outperform public corrections agencies rest
on two assumptions. The first is that there are significant differences between public and
private management, that business firms are necessarily more “efficient,” “effective,” and
“innovative” than government agencies, and that these advantages of private management
are universal: they obtain whether the task is picking up garbage or locking up prisoners.
The second assumption is that the public sector’s administrative experience in corrections
has been an unmitigated disaster: prisons and jails have been, and continue to be, horrible
places that are horribly run.

In my judgment, both assumptions are false. The former is the product of faulty logic,
poor conceptualization, and lazy moralizing; the latter results from an inadequate
appreciation of the fact that the performance of public agencies has varied enormously
and in ways that are clearly traceable to differences in public-management practices. Let me elaborate.

There is more than human caprice behind the fact that some tasks are in public rather than private hands; there are nontrivial reasons why we have both government and business, both politics and markets, both public agencies and private corporations, both MPAs and MBAs. In the public sector, the relationship between valued inputs (people, money) and desired outputs (less crime, better public health) is often unclear and may even be impossible to specify with any degree of precision; hence “efficiency”—i.e., maximizing output for a given set of inputs or minimizing the inputs needed to achieve a given level of output—is harder to measure. The political and legal constraints on what work gets done, how, and by whom tend to be far greater in the public arena. “Sophisticated” management theories and techniques may or may not help in the private sector, but they are almost always more likely to come away limping when applied to public tasks. Such verities about public versus private management have been discovered by successive generations of top-flight businessmen who have entered government service only to find that the lessons of Wall Street have limited application in government.

Nevertheless, suppose we wanted to make hard-and-fast evaluations of public versus private correctional management. What performance measures could we use? Most students of corrections would agree that there is little, if any, clear relationship between institutional penal practices, on the one hand, and rates of recidivism, on the other. Basically, prisons and jails yield an imprecise mix of incapacitation, deterrence, retribution, and rehabilitation—or so we tend to believe. It is quite difficult, however, to specify the relationship between any one of these ends and how penal facilities are managed. There are, however, at least three performance indices that may be less ambiguously related to managerial practices: order, amenity, and service.

By order, I mean the absence of individual or group misconduct that threatens the safety of others behind bars—simply stated, no assaults, rapes, riots, suicides, or murders. By amenity, I mean anything that is intended to increase the inmates’ comfort—clean living quarters, good food, color television sets. By service, I mean anything that is intended to enhance the inmates’ life prospects—programs in remedial reading, vocational training, work opportunities, and so on.

Measured accordingly, are all of America’s prisons and jails simply wretched, or are some of them relatively safe, humane, treatment-oriented, and productive? There are extremist views on both sides of this question, but the available data suggest that public penal institutions have been, and continue to be, a terribly mixed bag: some are clean, others filthy; some are orderly, others riotous; some offer many programs, others none at all; some cost over thirty thousand dollars a year per inmate, others under ten thousand. Certain public corrections agencies have improved over time; others have gotten worse; and still others have changed little, if at all. One thing, however, seems clear: the quality of life inside prisons and jails depends mainly on the quality of penal management. Through more humane and intelligent institutional management, prisons and jails can be improved, even when budgets are tight, facilities are dilapidated, and inmate populations are large and dangerous.

767
Draft — do not distribute!
One remarkable but untold correctional success story is the Federal Bureau of Prisons (BOP). Before the BOP was established by an Act of Congress and approved by President Hoover on May 14, 1930, inmates in federal custody were subject to the arbitrary and often cruel treatment of whichever of seven federal wardens they happened to be assigned to. Inmates ate slop served from buckets, and discipline was harsh. Under its first director, Sanford Bates (1930-1937), things began to improve. Under its second director, the dynamic James V. Bennett (1937-1964), the BOP made an incredible number of positive innovations in the care and custody of prisoners. The BOP’s third director, Myrl Alexander (1964-1970), consolidated these gains; its fourth director, Norman A. Carlson (1970-1987), instituted policy directives, recruitment and training programs, and scores of other changes that made the BOP’s facilities more safe, humane, and productive. While the BOP has had, and continues to have, its sore spots, its record over the last fifty-eight years provides solid evidence that good public correctional management is possible.

Like the BOP, other public corrections systems have had talented executives, conscientious workers, a prudent penal-management philosophy, favorable laws, political support, and good community relations.7 Under such administrative conditions, publicly managed prisons and jails have produced what most of us would probably consider acceptable (if far from ideal) levels of order, amenity, and service. The public sector has constructed, financed, and, most important, operated safe, clean, productive, and reasonably cost-effective institutions. The government has administered many prisons and jails at a reasonable human and financial cost.

In the absence of such administrative conditions, however, publicly managed systems have done less well. In some places, prisons and jails have been run by oppressive, uncaring officials; in other places, they have resembled states of nature, in which violent inmate predators rule; in still others, administrations have bounced between the poles of anarchy and tyranny, under conditions that judges have found unconstitutional and average citizens would find repulsively inhumane.

The crucial point, however, is that the administrative performance of public prisons and jails has not been uniformly bad, and there is no reason to suppose that the performance of private facilities would be uniformly good. To believe otherwise would be to believe that there is something magical about the private sector. In the world of the economics textbooks, a world of perfect competition and ceteris paribus assumptions, the market is magical. In this world, it is certain that efficient market-driven organizations can give us whatever bundle of goods and services we demand, from “widgets” to more complex commodities. But in the real world, organizations—whether public or private—tend to succeed when they combine good workers with sufficient resources under the right conditions, and to fail when they do not.

7 While none of them is (or was) perfect, some of my examples would include the California Department of Corrections under Richard A. McGee (1944-1967); the Texas Department of Corrections under Dr. George Beto (1962-1972); the South Carolina Department of Corrections under William Leeke (1968-1987); and the New Jersey Department of Corrections under William Fauver (1973-present).
The core administrative issue here is not public versus private correctional management, but the possibility of competent correctional management. There has been, and continues to be, variance in the performance of public-sector corrections agencies. By probing the political, budgetary, and other conditions associated with the better institutions and programs, public managers can learn from each other’s mistakes and successes. Rather than abandoning governmental stewardship of prisons and jails, we ought to give public administration a chance.

**WHO SHOULD GOVERN PRISONS?**

For the sake of argument, however, let us suppose that, contrary to the foregoing discussion, corrections corporations can operate successfully on a wide scale, construction, financing, and managing everything from tiny new community centers to massive old maximum-security prisons. And for the sake of argument, let us concede, without any qualifications, that private firms can indeed maximize services while minimizing costs in an abuse- and corruption-free environment, satisfying all political, legal, and other constraints. In other words, let us grant for the sake of argument that private prisons and jails are eminently feasible. Must it then follow that they are desirable? Is the private operation of prisons and jails, however instrumental it may prove to be in reducing costs and bettering services, justifiable morally? Does the government’s responsibility to govern end at the prison gates? Who ought to administer justice behind bars?

Behind each of these questions are prior ones. For example, where (or how) does one draw the line between “contractually deputized” private individuals and “duly authorized” public authorities? Since the authority wielded by public administrators is delegated to them by “the people,” by what strange metaphysic does the delegation of that authority to private firms constitute any sort of a moral (or constitutional) problem? The fact that “many different officials contribute in many different ways to decisions and policies in the modern state” gives rise to what has been aptly labelled “the problem of many hands”—the problem of deciding who in government is responsible for political or policy outcomes. The private management of prisons and jails, however, brings into focus what may be termed “the problem of whose hands”—the problem of deciding whether the moral responsibility for given communal functions ought to be lodged mainly or solely in the hands of government authorities.

In my judgment, to remain legitimate and morally significant, the authority to govern behind bars, to deprive citizens of their liberty, to coerce (and even kill) them, must remain in the hands of government authorities. Regardless of which penological theory is in vogue, the message that “those who abuse liberty shall live without it” is the philosophical brick and mortar of every correctional facility. That message ought to be conveyed by the offended community of law-abiding citizens, through its public agent, to the incarcerated individual. The administration of prisons and jails involves the legally sanctioned exercise of coercion by some citizens over others. This coercion is exercised in the name of the offended public. The badge of the arresting policeman, the robes of the
judge, and the state patch on the uniform of the corrections officer are symbols of the inherently public nature of crime and punishment.\textsuperscript{9}

\textbf{RENT-A-PRESIDENT?}

The moral implications of privatizing the administration of this central communal function—administering justice for acts against the public welfare—can be felt by entertaining morally analogous situations. Suppose, for example, that the issue were not who ought to administer justice to the community’s offenders but who ought to administer rewards to the community’s heroes. Consider the following hypothetical scenario. You have worked tirelessly on behalf of your fellow citizens. You have discovered a cure for cancer, or negotiated successfully with our foes abroad, or made a major contribution to the performing arts. For your good deeds, you are to receive a National Medal of Honor. The big day arrives. The crowd is assembled on the White House lawn; the Marine Band begins to play. You gaze into the crowd and notice that the distinguished-looking guests (and the “Marine” musicians) have little pins on their lapels that read “MCA” (for Medals Corporation of America). The ceremony is grand. The crowd roars as your name is called and the “president” (better looking than the real commander-in-chief) shakes your hand and embraces you. You know for a fact that by every tangible measure—the physical quality of your medal, the warmth of the presenter, the duration and intensity of the crowd’s ovation, and the sound of the music—MCA gave you a ceremony that was far superior to what you would have received had the government and its officials presided. Moreover, MCA spent only one-third of what the government would have spent. Are you satisfied?

Or imagine that a private consulting firm consisting of the nation’s wisest, most seasoned, and most widely respected statesmen could do a superb job of selecting our next president, at a tiny fraction of what it now costs to stage primaries and a general election. The very thought send a moral chill up my spine because I believe, as most Americans no doubt do, that the public, democratic nature of the process of presidential selection has a value independent of the outcome. Thus, in a situation where it was certain that Candidate X was going to be selected regardless of whether the choice was delegated to the presidential head-hunting firm or left in the hands of the voters, I would not hesitate to choose the public process even though it was far more expensive. Indeed, I would probably do the same even if the “private presidents” corporation would select my favorite candidate while the voters would not.

Or suppose that CCA has made it really big. They have proven that they can do everything the privatizers have promised and more. The corporation decides to branch out. The company changes its name to CJCA—Criminal Justice Corporation of America.

\textsuperscript{9} That these symbols are not empty is presupposed in a wide body of public laws. For example, the authors of a thorough 1977 survey of private policing in the United States observed that both “private citizens and private security personnel are generally prohibited from wearing badges or uniforms that might be confused with those worn by public police.” Some states (for example, California) go well beyond simple regulations proscribing impersonation, forbidding private police from doing anything that might “create the impression that they are acting under government authority”; other states (New York, for example) prohibit private investigators from “owning, exhibiting, or displaying a shield or badge in the performance of his duties.”
It provides a full range of criminal-justice services: “cops, courts, and corrections.” In an unguarded moment, a CJCA official boasts that “our firm can arrest ’em, try ’em, lock ’em up, and, if need be, fry ’em for less.” Is there anything wrong with CJCA?

In each case, the relevant human collectivity can be said to have abdicated its duty to reward, choose among, or punish its members—the nation at large its heroes, the electorate its leaders, the community of law-abiding citizens its criminals. Implicit in the “farming out” of such responsibility is a denial of the group’s moral integrity. Ought not the nation to express its gratitude directly through the agency of its elected and duly appointed public representatives? Is it not the voters alone who may bestow authority and legitimacy on occupants of the Oval Office? Is it not the community of law-abiding citizens that is offended by criminals and should dole out their punishment?

**The Moral Writ of the Community**

If every tangible advantage clearly belongs to private management, does it matter whether the corrections officer’s patch reads “Corrections Corporation of America” or “State of Tennessee”? Where the governing of prisons is concerned, is management by private hands morally distinguishable from management by public hands? At a minimum, it can be said (both in theory and in practice) that the formulation and administration of criminal laws by recognized public authorities is one of the liberal state’s most central and historic functions; indeed, in some formulations it is the liberal state’s *raison d’être*. In the opening chapter of Locke’s *Second Treatise of Government*, political power itself is defined as “a right of making laws with penalties of death, and consequently all less penalties, for the regulating and preserving of property, and of employing the force of the community, in the execution of such laws . . . , and all this only for the public good.”

Criminal law is the one area in which Americans have conceded to the state an almost unqualified right to act in the name of the polity, and hence one of the few places in which one can discern an American conception of political community that is not a mere collage of individual preferences. It is not unreasonable to suggest that “employing the force of the community” via private penal management undermines the moral writ of the community itself.

But what about the argument that the community’s moral writ remains intact as long as the pre-sentencing process remains in public hands? Returning to our earlier illustration, one might assert that CJCA would be wrong, but CCA is not. There are at least two problems with this position. First, it rests on the untenable presumption that the administration of penal facilities involves little or no exercise of discretion by the administrators, or at least none that would affect the duration of an inmate’s stay or the basic conditions of his confinement. There is a mountain of empirical studies that show how much discretion at every level—from the commissioner’s office to the cell block—is of necessity vested in those who run prisons and jails. Any normative case for private penal management that hinges upon a resurrection and acceptance of the discredited notion of a politics-administration dichotomy is prima facie too weak to require a rebuttal.
Second, it is simply unclear how one can distinguish morally between private and public courts, and between private and public policing, and yet see no moral difference between private and public corrections. The logic required to embrace a CCA while rejecting a CJCA stems from a species of moral confusion of the same sort that underlies discussions of the death penalty in which supporters of privatization declare that while corporations may “play jailer” (even for life), only the government should “play executioner.” Those who claim that a CCA is morally legitimate have little basis for rejecting not only private police but also private judges, juries, and executioners.

One could imagine the staff of a private corrections firm being put through some sort of “publicization” ritual—raising their right hands, reading the state constitution, saluting the governor, and so forth. But as long as their organization remained private, the chain of delegated authority could be said to have a rotten link.

There is no limit to the number of hair-splitting dimensions that can occur on such points. Though I have only scratched its surface, enough, I hope, has been said to suggest that serious moral problems surround (and may weigh against) the private management of prisons and jails.

A BETTER ALTERNATIVE

Privatization efforts have been fueled by the belief that public correctional institutions are too crowded; but the crowding problem is less acute than is commonly supposed. Privatization initiatives have been offered as ways of tightening the reins on galloping correctional budgets, but the public sector has made innovations that promise the reduce the corrections tax bill. Privatization ventures are driven by the perception that public correctional managers have failed; but the public record is by no means unrelievedly bleak, and in some respects it is quite impressive. The problems of crowding, rising costs, and failed management are most evident in the area of high-security prisons and jails, but at present the privatizers offer no help for these institutions.

When privatization proponents assert that the firms will do better because they, unlike their public-sector counterparts, will be immune to the political, administrative, and financial woes of government’s “red tape,” they should be reminded that “one person’s ‘red tape’ may be another’s treasured safeguard.” Even a cursory review of the historical, political, and administrative issues surrounding private prison and jail management raise grave doubts, and not a few fears, about the prospects of privatization in this area.

The central moral issues surrounding private prison and jail management have little to do with the profit motive of the privatizers and much to do with the propriety, in a liberal constitutional regime, of delegating the authority to administer criminal justice to nonpublic individuals and groups. For much of American history, government has allowed too many of the community’s prisons and jails to be ill-managed, under-managed, or not managed at all. Especially in light of the progress that has been made over the last two decades, no self-respecting constitutional government should again abdicate its responsibility to protect and guide criminals in state custody.
The most promising corrections alternative is to get down to the nitty-gritty business of governing prisons well with the human and financial resources available in the public sector, jurisdiction by jurisdiction. Where public corrections agencies have managed their discretion and coercive powers with common sense and compassion, prisons and jails have been relatively safe, clean, and cost-effective. But far too many institutions remain unsatisfactory, and more progress in public correctional management can and must be made.

While private corrections firms like CCA have run admirable juvenile centers and other facilities, we are most likely to improve our country’s prisons and jails if we approach them not as a private enterprise to be administered in the pursuit of profit, but as a public trust to be administered on behalf of the community and in the name of civility and justice. The choice is between the uncertain promises of privatization and the unfulfilled duty to govern.


I. INTRODUCTION

The evolution of the contemporary criminal justice system can be described in terms of two distinct shifts in the relationship between state and society and in the role of law in mediating that relationship. The more obvious of these shifts is the direct expansion of the modern state’s sovereignty over the affairs of citizens, as is perhaps best evident in the demise of libertarian interpretations of the Fourth Amendment and attendant increases in state prerogative. The second, less obvious shift involves a dramatic erosion of the state’s monopoly of criminal justice functions—that is, the “privatization” of criminal justice. Although evident in the rise of victims’ rights regimes, private police forces, and even private financing of criminal prosecutions, this privatization of criminal justice may be most notable in the re-emergence of the private prison.

At present, about 120,000 men, women, and children are incarcerated in privately managed, for-profit jails, prisons, and detention facilities—in the hands, literally, of an oligopolistic industry with annual revenues of at least $1 billion and perhaps far more than this. Though these figures are still dwarfed by the two million jail and prison inmates in America today and the huge economic scale of incarceration generally, they also reflect an annual rate of growth in private incarceration that has over the past several years averaged four times that of the (already astronomical) overall growth in criminal incarceration. Perhaps more tellingly, this industry only emerged in its modern form about fifteen years ago.

The straightforward expansion of state sovereignty in the criminal context has garnered substantial criticism from practical, political, and jurisprudential angles. But the
privatization—or, as one should say, the semi-privatization—of criminal justice has escaped comprehensive criticism. Instead, critiques of the private prison tend to focus narrowly on the institution’s practical, legal, or general normative failures, and do so to the exclusion of any sustained focus on the private prison’s implications for the changing relationship between state and society and the way the law mediates that relationship. In short, there persists a crucial failure to mount a jurisprudential critique of the private prison.

In this Article, I argue that a rigorous jurisprudential critique of the private prison shows that the private prison tends to distort dramatically the relationship between state and society in the criminal context, and does so in a way that contradicts the most central of liberal legal precepts: the rule of law. Such a rule of law critique of the private prison sees in that institution a key development: the simultaneous expansion and diffusion of state sovereignty, accompanied by the thorough merger, or interpenetration, of public and private realms. This process renders the private prison utterly inconsistent with the rule of law’s central aspiration: the restraint of sovereignty and the concomitant realization of negative freedoms and minimal equality by the mediation of law. From this perspective, and following Stanley Cohen’s more general critique of criminal justice reforms, I argue that the private prison inevitably constitutes an extravagant, but at the same time insidious, aggregation of state power in a context where such power is being deployed in a largely irrational way. Such a critique reveals with considerable irony that the privatization of prisons, a movement ostensibly based on the ideal of the “minimalist state,” is actually the antithesis of such an ideal.

This notion that the rule of law is premised on an aspiration to restrain sovereignty is shared in key respects by such diverse figures as the liberal, Friedrich von Hayek, and the neo-Marxist jurist, Franz Neumann. The rule of law’s antithesis to the private prison follows more specifically from the idea that the rule of law’s sovereignty-restraining aspiration presupposes the clear demarcation of the sovereign and the transparency of sovereignty, which in turn presupposes the substantial segregation of public and private realms. Although these conditions have never been fully realized in any society, and in many ways constitute a problematic formulation, the regime implied by the rule of law retains a contingent normative value, guaranteeing a baseline of liberty and equality. Inasmuch as the private prison is premised intrinsically on the simultaneous extension and diffusion of sovereignty, and also on the merger of the public and private, it is intrinsically at odds with the rule of law and emerges as a fundamentally illiberal development—far more problematic than the public prison already is.

Of course, the doctrine of the rule of law is not a rule of law in the literal sense: a “violation” of the rule of law does not make something illegal in any positive sense. Still, this Article’s critique of the private prison is practically, and perhaps even legally, relevant in several ways. First, it gives general jurisprudential structure to a debate that largely lacks such structure. Second—and perhaps this is part of the basis of its claim to be jurisprudential—this Article’s critique of the private prison describes a common link between erstwhile separate normative, practical, and legal critiques of this institution. This Article argues that there are very concrete connections between the private prison’s anti-rule of law character and its practical, legal problems. The history of the private
prison’s antecedents, especially of the convict lease system, shows clearly that the private prison’s ubiquitous tendencies to corruption, legal ambiguity, and the augmentation of state power are tied organically to its inconsistency with the rule of law. In other words, history shows that the otherwise abstract antithesis between the private prison and the rule of law seems to be accompanied by inherent tendencies to translate into significant practical and legal problems as well.

In order to develop these arguments, this Article proposes in Part II a definition of the rule of law that transcends conventional ideological positions, that is centered on the concept’s sovereignty-restraining aspiration, and that stresses the doctrine’s contingent rationality. Such a definition is shown to imply the separation of public and private and the clear demarcation of sovereignty. Here I begin to develop the notion that the rule of law is fundamentally inconsistent with the private prison. Part III tackles the crucial and deceptively complex matter of defining the private prison. Here, I outline the curious history of the private prison and show that the private prison, in some form or another, is the historical norm. . . . In Part V, I describe in more detail the characteristics and origins of the contemporary private prison and expose the shortcomings of existing attempts to critique the phenomenon. This Part also comprises the main critique of the private prison from a rule of law perspective. Here as well I consider the limits of this Article’s critique with regard to other modes of privatization. Finally, Part VI offers a cautionary conclusion that considers the implications of the private prison in broader context.

II. RULE OF LAW AND THE RESTRAINT OF SOVEREIGNTY

The question that always greets any scholarly mention of the rule of law is whether the concept actually possesses any real scholarly value, or whether it is merely a vehicle for partisan rhetoric. Certainly the rule of law enjoys a prominent place in contemporary legal, political, and social discourse. A reader need only peruse, for example, contemporary finance magazines, human rights tracts, and international politics journals to encounter a steady stream of arguments and claims couched in rule of law terms. But more often than not, the rule of law operates in these fora as a fuzzy euphemism for a set of institutions that (quite conspicuously) seems to approximate the aspirations and self-images of contemporary Anglo-American society. In contemporary discourse, the rule of law is variously equated with the legal structure of so-called free-market finance capitalism, the norms of the western human rights agenda, or a constitutional order that usually is similar to the American model.

While even semi-serious appeals to the rule of law provide some evidence that the concept has not been completely trivialized, there is something rather troubling, or at least grossly inadequate, about the usual manner by which the rule of law is construed. Contemporary perspectives on the rule of law do little justice to the jurisprudential character of the concept. Even in scholarly circles, the concept usually is reduced to its bare functional or instrumental aspects, rhetoric, slogans, and operational figments. Such a move is problematic not simply because the rule of law can mean so much more, but also because the jurisprudential themes skipped over by this logic contain the concept’s most fundamental, critical components.

775
Draft — do not distribute!
At the logical and genetic core of the rule of law is a fundamentally jurisprudential agenda: the restraint of sovereignty by law and the concomitant idea that in a rational society, law aggressively mediates the relationship between the state on the one hand, and civil and domestic realms, on the other. It is this sovereignty-restraining aspiration that underlies and really accounts for the rule of law’s more specific and formal claims: generality, universality, and non-retroactivity of legal norms, separation of powers, and so forth. The present critique of the private prison is jurisprudential because its rule of law critique of the private prison is couched squarely in such sovereignty-restraining logic.

This sovereignty-restraining notion of the rule of law surfaces in the earliest attempt to define the concept. In the Western world, the idea of the supremacy of law over politics emerged in classical society and was entrenched substantially (albeit with limited scope) by the Middle Ages. According to Geoffrey Walker, the rule of law, at least in concept, survived the oscillations of royal powers in the Medieval and early modern eras to emerge in modern times as a viable set of claims against the unmeasured administration of power by sovereigns. But in the final analysis it is only with modern interpretations of the rule of law—at least the serious ones—that a consistent emphasis is placed on the rule of law’s sovereignty-limiting character. Montesquieu, especially, offers the notion that the rule of law demands the comprehensive restraint of sovereignty. In such a mode, the concept of the rule of law was eventually able to comprise the juridical ideology, one might say, of bourgeois ascendancy, forming a set of effective challenges to aristocratic prerogative as well as the juridical basis for “free” competition within the structures of the capitalist market.

An essentially sovereignty-restraining view of the rule of law to a large degree transcends the political perspectives of its modern exponents. Thus, such a view pervades the work of the archliberal Friedrich von Hayek and the common law parochialist A.V. Dicey, the consummate liberal centrist Max Weber, as well as the Marxist jurists Franz Neumann and Otto Kirchheimer, and the Marxist historian E.P. Thompson.

... For serious advocates of the rule of law, the concept’s sovereignty-limiting function is absolutely key. How, one might wonder, does the rule of law accomplish such an end? On this point, too, there actually is substantial consistency. Serious defenders of the rule of law insist, in the first instance, that the concept, while not a statement of law in its own right, must be understood to possess a minimal amount of legal authority, to operate, as Hayek puts it, as a “meta-legal doctrine,” or as Neumann implies, as a secular, quasi-natural law doctrine. Beyond this, there is agreement too that the rule of law’s sovereignty-restraining function rests on the application of a set of subsidiary principles—in particular: generality, neutrality, universality, non-retroactivity, separation of powers, the insularity of the legal system, and so forth—to the normative structure of the law.

Such an agenda implies a social system premised on the segregation of political and legal authority, formal equality, and “negative” freedoms vis-à-vis the sovereign. This relationship in turn presupposes a mutually exclusive, but at the same time complementary, distinction between the public realm, the home of sovereignty, and the
private realm, the negative reflection of sovereignty. The resulting regime is not unproblematic. For example, the public-private distinction tends to sanction the substantial residual sovereignty (in the form of tyranny, really) and inequality of the domestic realm. And as both Marxists and Weberians recognize, the rule of law simultaneously facilitates and reflects the exploitative, alienating dimensions of capitalist civil society. Nevertheless, like Marx, Neumann and Kirchheimer, and, to some extent, Weber, I think it is important to embrace the rule of law as the apogee of legal rationalization under our existing historical horizon. Even more fundamentally, it may be that some variant of the rule of law will always be preferable to a system in which power knows no restraint; it may be that the rule of law is essential to any rational social order.

Of course, the key question is then, how does the rule of law relate to the private prison? A number of authorities have emphasized the relevance of rule of law principles in rationalizing the criminal justice system in general. The rule of law is understood as a basis for the critique of discretion, inequality, and unbounded expressions of state power in the institutions of criminal law and criminal procedure. In fact, some commentators contend that the rule of law directly forecloses private criminal justice functions. It is indeed possible to say that the rule of law, because it implies the sovereignty of law, vitiates private justice simply because private persons neither can construct nor implement between themselves general, formally equal, predictable, and non-retroactive legal norms. It also is possible to argue as well that private law-giving is inconsistent with the insular, self-contained pretensions of the rule of law. There are indeed many ways to draw out antitheses between the rule of law and the privatization of law, especially in the criminal context.

Yet it is in the most fundamental way, I think, that the rule of law speaks critically to the private prison. My main thesis is as follows: The rule of law evidences an essential antipathy to sovereignty and a concomitant ambition to restrain sovereign deeds with subsidiary norms like generality, universality, separation of powers and so forth. The aspiration behind these norms only can be realized if the sovereign is, in the first place, a legally and politically transparent entity with clearly demarcated boundaries. The idea of freedom from sovereignty that the rule of law claims, requires that the sovereign have definite limits, that when an institution or person acts, we can know clearly if it or she is the sovereign. Who, otherwise, is to be restrained from whom? It is in this manner that rule of law norms presuppose the clear segregation of state from civil and domestic society and of public from private realms.

It is from such a perspective that Hannah Arendt, in her otherwise problematic classic, quite accurately defines the absence of the rule of law as a signature aspect of twentieth century totalitarianism. Among others, Neumann and Kirchheimer also note how the abrogation of the rule of law provides the legal foundation of fascism; how, for example, the fascist negation of rule of law norms authorized the erosion of legal generality and the public-private distinction, and in turn facilitated a massive interpenetration of public and private realms that featured the state’s domination of private life, the frequent resort to individualized and retroactive laws, the complete politicization of legal process, and the domination of the state by private cliques and quasi-public political parties.
Needless to say, the return of the private prison itself does not necessarily imply the advent of totalitarianism or fascism or any other kind of far-reaching reconstruction of our political and legal universe. But inasmuch as the prison is in many respects the quintessence of a state’s sovereign function, and inasmuch as the private prison so thoroughly merges the private and the public and blurs the boundaries of the sovereign, the private prison cannot help but be antithetical to the rule of law. Immediately, this exposes the private prison as fundamentally problematic on at least an abstract, normative plane. Perhaps more critically, I shall argue that the private prison’s concrete problems have roots in a juridical structure built around the abrogation of the rule of law.

III. LOCATING THE PRIVATE PRISON IN PRACTICE AND IN HISTORY

In order to critique the private prison it is, of course, necessary to define the institution. Neither the term “prison” nor “private prison” has a self-evident meaning. Even outside of its many metaphorical usages, the term prison has long been construed in diverse ways. At various points it has contemplated everything from facilities for detaining juveniles and undocumented immigrants, to “halfway” houses, to city and county jails housing misdemeanants and those awaiting trial, to the quintessential “big houses,” huge self-contained edifices brimming over with hardened felons, that continue to dominate the prison landscape.

A. Conceptualizing the Private Prison

It is quite difficult to say what is a prison and what is not. It is surely impossible to define the prison in a logically or nominally incontestable fashion. This problem is underscored, it seems, not only by the frequency with which institutions that are, in one respect or another, non-criminal, non-custodial, non-punitive, or not closely related to the state are nonetheless described as prisons, but also by the tangled history of the development of the prison. For legal scholars and social scientists in particular, the only way to box the meaning of the prison is in effect to conceive of it through its function, which is to say, to construe the prison as an institution that, for example, involuntarily confines persons committed by the state to relatively long terms of incarceration (e.g., one year), for violating the public, criminal laws. For the most part, and not least for reasons of consistency, the present inquiry adheres to this definition. But much more important to this Article’s critique than any such conceptual details is the idea of prisons, including private prisons, as fundamentally coercive and implicitly violent places where the sovereignty of the state over its citizens, as a prerogative of total control, assumes its most extreme form (with the probable exceptions of capital punishment and war-time compulsory military service) and where the hands of the state are always evident. From this perspective a whole range of prison-like facilities looms equally problematic vis-à-vis the rule of law, in kind if not also degree.

What makes a prison “private”? Michel Foucault famously defines prisons as intrinsically “private” in the sense that they secret from view punishment practices that were heretofore “public” in a parallel sense. Whatever the merits of this formulation, this Article is concerned with private prisons in the proprietary sense, as institutions that are managed and sometimes owned by non-state entities. Beyond this, though, the meaning
of the private prison remains no more self-evident than that of the prison as such. The limited character of the state in American society complicates this issue. To the extent that the state is not ubiquitous, and that the prison is not entirely hermetic, some aspects of every prison are always private. From the labor of its employees, to provisions for inmates’ subsistence needs, to the land and capital that comprise the prison’s physical structure, each exemplifies every prison’s endemically, if partially, private character. In this sense, it is only possible to imagine a fully public prison either in a thoroughly totalitarian society or when the prison itself is (and this would negate its quality as a prison) an entirely self-contained society.

At the same time, no prison in the contemporary world can be fully private. Every prison remains intimately connected to the state, incarcerating inmates arrested, prosecuted, and sentenced by the state for violating the (still) very public criminal laws and their analogues (for example, juvenile offender laws). In this sense, the privatization of prisons is much unlike, say, the privatization of steel mills or utilities or even schools, which may be mandatory and relatively coercive in operation, but to a much more limited degree than prisons. Another dynamic that keeps the private prison very public is that private prisons operate exclusively on revenues derived from the state.

The inherently public aspect of private prisons is a crucial point. It is a principal argument of this Article that in the final analysis the private prison comprises the extension and diffusion—and in no way the negation or displacement—of the sovereignty of the state. As we shall see, this endemic confusion of public and private ushers the private prison into a state of inevitable illegitimacy.

As a practical matter, both critics and proponents of the private prison tend to define it as an institution for criminal confinement that is wholly managed and operated by a private firm. Such a perspective excludes not only the typical public institutions, but also those that are provisioned in a more extensive but still piecemeal way by private entities (for example, that out-contract food or laundry services), as well as institutions whose physical structures are privately owned but are nonetheless publicly managed or operated. As with the matter of the defining the prison, I again accept the impossibility of drawing clear, logically incontestable lines around the private prison. As the following analysis makes clear, for this argument it is the degree of private involvement that makes the private prison so problematic—and not any particular qualitative aspect.

**B. The Private Prison in Historical Perspective**

Despite its contemporary ubiquity, the prison is a relatively recent fixture in Western (indeed, every) society. Moreover, from the outset the prison was infused with private ownership and control, and with private functions, in many respects quite similar to the contemporary private prison. Only in the nineteenth century did the prison come to constitute a common mode of criminal sanction in the United States, and only in the twentieth century did the prison come to comprise a primarily publicly managed affair. It is fair to say that the prison was private long before and long after it was, in fact, a prison. This history does much to anticipate the character of the contemporary private prison, its juridical structure and its dysfunctions.
The historical development of the prison is utterly steeped in the interpenetration of the public and private realms in Western society. Although the distinction of public and private realms was long ago introduced to the Western world in a very furtive way—for example, appearing in the attempt of the early Roman law to distinguish public and private wrongs—as a practical, concrete matter the distinction only really took hold in the nineteenth century, and only then (as countless realists and critics have indicated) in a most incomplete fashion. Pre-modern societies, especially ancient ones, are rather uniformly characterized by the confusion of public and private realms and, where law itself had attained articulate form, a confusion of public and private legal norms.

Frequently enough, the underdevelopment of the public-private distinction manifested itself in the juxtaposition of civil and criminal legal regimes and, perhaps even more saliently, in the juxtaposition of public and private control of the administration of “criminal” sanctions. Only recently in Western society did there exist anything even approaching a public monopoly of criminal justice functions. Of course, state-structured punishments did appear in the pre-modern world, including the European world. But only rarely did this involve any kind of formal, punitive incarceration, which almost always was used for purposes of criminal and civil detention, and not for punishment as such. Much more typical were extra-legal punishments or punishments based on the application of fines and tort-like sanctions, forced labor, banishment and exile, corporeal punishment, and the like. These practices were consistent not only with the barbarism of the day but, more importantly, with existing structural and material realities: the rigid social relations, the absolute lack of social surplus, and the general shortage of labor in such societies. In this kind of historical context the prison as we know it—and as Georg Rusche and Otto Kirchheimer stress—could not assume a central place in the system of social control.

When the practice of punishment by incarceration did appear in pre-modern society, it tended to reflect within itself the prevailing confusion of the public and private in society as a whole. Almost always, early prisons, which in scale, function, and internal structure were more like contemporary jails than anything, either were privately owned or managed, or served transparently private functions, or both. From its very inception in Western society, the prison was used to achieve such private ends as the collection of civil debts, the punishment and secreting away of rivals, and the administration of domestic tyranny. In medieval Europe, this tradition played out perhaps most conspicuously in the punitive use of prisons to maintain order within the essentially private domains of noblemen and clergy. In the early modern era, this dynamic prevailed in the use of prisons to detain upper-class delinquents and the insane.

To an equal degree, the early prison almost always was a privately owned or managed affair. Feudal manors maintained prisons that were private in the truest sense: privately functioning, privately managed, and privately owned. From medieval times through the Industrial Revolution, the maintenance of European jails tended to be the personal responsibility of local sheriffs and their analogues, enterprising minor noblemen, or everyday entrepreneurial “keepers.” Not just an obligation, though, this function remained a “business proposition” until at least the end of the eighteenth century, with the keepers and “franchisees” taking fees from the state and inmates (or “customers”) alike. In the sixteenth century in Europe, there arose an institution that completely merged private (or at least non-criminal) functions with for-profit management: the so-
called “house of correction.” The house of correction, which united under private management the functions of poorhouse, jail, and manufactuary, also juxtaposed under private management very minimal public safety functions (as we would now think of them) with public welfare and labor-control functions.

The private, labor-exploiting character of the house of corrections was not at all unique among early modern punishments. Convict “transportation,” especially, closely replicated the early prison’s thorough juxtaposition of public and private characteristics. Transportation saw the forced removal of tens of thousands of “criminals” (and the term has to be used advisedly) from Europe to places like Australia, North America, and New Caledonia, where they spent their terms laboring for private contractors, often in the immediate custody of such contractors. Other punishment regimes of this time, forced labor on public works and galley slavery, were not exactly private; yet they too operated on a statist, mercantilist logic of public profit-making that once again underscored the confusion of public and private in these times and places.

In the United States, the confusion of public and private impressed itself into the history of criminal justice generally, and that of the prison more specifically. In colonial America, there were no true prisons and very few jails to start with, and incarceration was rarely resorted to as formal criminal punishment. As in medieval Europe, labor was too precious, social surpluses too short, and culture too backwards to justify such institutions. Punitive incarceration was not, of course, unknown. However, as elsewhere, it was subordinate to other practices like banishment, fines, corporeal punishment, and, of course, outright vigilantism. And where there were jails and prisons, they were typically privately run operations—especially, it seems, in that even ostensibly public officials derived their salaries both from fees as well as from charges levied on inmates. As in Europe, these facilities often took the form of privately-run houses of corrections.

America generally is regarded as the birthplace of the modern prison—the prison, that is, as a place for large-scale, long-term, and punitive incarceration. Indeed, the foundations of the modern prison were set with the inauguration of two rather distinct models of the “penitentiary”: the so-called “Walnut Street” or “Pennsylvania” system, established in the 1780s; and the “Auburn” or “congregate” system, which grew out of the Walnut Street system early in the nineteenth century. Both of these systems were premised on a set of common practices, centered around the segregation of inmates into private cells and the enforcement of stringent rules of inmate silence. The main difference between the two lay in their distinct ways of mobilizing inmate labor. The Walnut Street system, which mandated silence by absolute solitude among inmates, lent itself only to the minimal, inefficient use of labor (for example, in a manufactuary mode). The Auburn system retained for a time the Walnut Street system’s rules of silence, but eschewed the institution of absolute solitude in the course of instituting a much more efficient, collective, factory-like system of labor. The Auburn system thus gave rise to a set of practices—most notably the mainly nocturnal use of private or semi-private cells combined with some type of collective labor in the day—which continue to inform the structure of contemporary prisons worldwide.
Though still infused with some private characteristics, each model of the penitentiary marked a significant movement towards the modern, public prison. Both systems also brought about the segregation of criminal inmates from civil detainees (for example, debtors). The inefficient Walnut Street system proved especially incompatible with the profitable employment of inmates, and hence with private management or ownership structures. And while the Auburn system could profitably employ labor, it was only rarely that this involved direct private control. Sometimes jurisdictions leased out inmates to contractors who installed their capital in the prison and directly managed the inmates’ work; more often this system involved a more subtle reduction of erstwhile public institutions to the logic of private market actors. Under the dominant “piece-price” system, prison administrators assumed the role of factory managers, organizing the (hopefully) profitable production of commodities by their inmates. The quintessential big house prisons employed this system to quite profitable ends, selling on the commercial market everything from shoes to furniture until the twentieth century.

At the same time that the rise of the penitentiary marked a move away from the private prison, other institutions from roughly the same period show just the opposite movement. Most obvious in this regard are the punitive aspects of Southern slavery. In many ways, slave plantations in the antebellum South were private prisons unto themselves. Masters and mistresses retained wide authority by custom and law to discipline their “properties” for all manner of deed, criminal or otherwise. Of course, slave owners did not entirely monopolize criminal justice functions with respect to their slaves, but they nearly did so, and the exceptions—for example, the modest and usually ineffective legal limits on owners’ disciplinary prerogatives—mainly prove the rule. In such a social system, incarceration exacted direct costs in lost labor, fines were utterly inapplicable, and public punishments of all sorts implied a loss of slaves’ services and an abrogation of owner’s prerogative. Accordingly, the punishment of slaves most often took the form of corporeal punishment, extra work, and punitive resort to the slave auction, all privately administered.

In the postbellum South, until at least the 1940s, the practice of substituting the plantation for the penitentiary continued in the guise of several different forced labor regimes. Among the more salient examples of these was the so-called “criminal-surety” system, whereby “criminal” offenders (for example, those who were “convicted” of violating very dubious vagrancy, petty larceny, or trespassing laws) typically were afforded the “opportunity” to exchange future labor for payment of their fines and “court costs” by local landowners and capitalists. Sometimes formal, statutorily authorized outcome of convictions, sometimes informal and the result of pretrial plea-bargaining, the result was always the same: reduction to peonage as criminal punishment (or, more accurately, criminal sanction as a source of peonage). The offender usually was bound to remain and labor at his surety’s establishment for months or even years earning minimal wages against his debt. . . .

[In Part IV, White discusses the convict lease system.]
V. THE CONTEMPORARY PRIVATE PRISON AND THE SOVEREIGNTY QUESTION

A. Conventional Critiques of the Private Prison

... From its resurgence in the 1980s, the private prison has been subjected to a number of more or less scholarly critiques. These critiques, which of course transcend the critique of criminal punishment as such, can be grouped in three overlapping categories: (1) those that focus on the private prison’s practical or “performance” shortcoming; (2) those that point to the private prison’s legally problematic characteristics; and (3) those that deem private prisons inherently problematic on a normative, usually moral or ethical, plane. A review of these perspectives reveals that, while each is often quite useful, they remain inadequate to a thorough critique of the private prison.

Practical criticisms of the private prison center on questions of performance—that is, on the ability of these prisons to achieve, for example, greater efficiency than their public counterparts or otherwise to provide adequate services measured in levels of violence and abuse, recidivism, and fiscal costs and relative to public prisons. Perhaps the most common variant of this critique focuses on a proposed contradiction between the provision of quality services of any kind and the financial self-interests of private prison contractors. Though many performance issues remain unresolved, there is more than ample evidence that private prisons are not dramatically more efficient than public institutions and have struggled to provide services even equal to that of public prisons. Aside from the underdeveloped state of the debate, the only thing consistently problematic about such critiques is that, whatever the focus, they consistently fail to connect narrow failings to fundamental jurisprudential problems.

Much the same can be said of typical legal critiques of the private prison. These critiques frequently focus on whether the private prison violates constitutional constructs like the so-called non-delegation doctrine, the due process clauses, or the Thirteenth Amendment. In other cases, critics confront the private prison’s relationship to traditional liability doctrines, both state and federal, statutory and common law. Indeed, the 1997 Supreme Court decision, Richardson v. McKnight, which squarely confronted the question of private prison guards’s entitlement to the privilege of qualified immunity in the context of § 1983 actions, apparently accounts for the majority of recent legal commentaries on this topic. Other items of concern in this area include the matter of access to private prison records and the legal issues engendered by intervening- or cross-jurisdictional private prisons. Again, the problem with such critiques is not an intrinsic one. As we shall see, the recurrence and relative insolubility of liability questions in the private prison context is an important premise of my own critique. The problem is rather a failure to connect narrow failings to fundamental jurisprudential problems.

More jurisprudential in spirit, if not in execution, are the numerous attempts to paint the private prison as something that is inherently obnoxious on general normative grounds. As often as not, the notion that private prisons are intrinsically improper appears in journalistic, usually editorial accounts. These efforts are remarkable both for their great conviction and, unfortunately, for their persistent failure to engage legal and therefore jurisprudential issues. The same is true of more sophisticated, scholarly efforts. For
example, such arguments by neo-conservative political theorist and penologist John DiIulio (the Princeton scholar who introduced the term “super-predator” to criminological discourse) against privatization are quite intricate and provocative and appear to have wielded some influence, at least in academic circles. Nonetheless, DiIulio’s arguments are couched in moral claims that he rigorously separates from practical and legal questions. For this reason, DiIulio’s work, too, remains largely inconsistent with a truly effective jurisprudential approach. Inadequacies of the same kind characterize the few “ethical” critiques that have been raised in this literature, including a notable effort by Richard Lippke, as well as the American Bar Association’s policy statement against private prisons. Of course, in my view the private prison is normatively problematic. But once again, I think it deficient not to connect normative claims to objective problems.

In sum, the main problems with the existing critiques of the private prison consist of a failure to deal with the juridical implications of the private prison and thus a failure to couch the private prison’s implications in a discussion about the way that the law mediates, or fails to mediate, the relationship between state and society. Likely, it is a consequence of these shortcomings that the existing literature is so fractured, that normative critiques remain disconnected from practical and legal critiques, while practical and legal critiques are disconnected from normative failures. For, as I have mentioned, it is the essence of jurisprudential discourse to draw out the connections among these issues, to see, for example, in the connection between practical and legal dysfunctions and erstwhile abstract normative failings. To eschew jurisprudential critique is therefore to guarantee not only that important questions about the fate of sovereignty are not discussed in the context of the private prison, but also that the fractured character of the debate is not mended into a comprehensive critique.

**B. The Contemporary Private Prison and the Perversion of Sovereignty**

The key to a thoroughly critical understanding of the private prison lies with its relationship to sovereignty. The private prison represents neither the straightforward retreat of sovereignty, nor its outright expansion. Rather, the private prison is fundamentally premised on a dynamic that combines these tendencies, that seems to represent both the apparent retreat and the advance of the state in the prison context. It is in this sense that private prisons must be understood in terms of the extension and diffusion of sovereignty.

Commenting broadly on criminal justice issues, Stanley Cohen argues forcefully that the coerciveness of the state consistently is expanded by apparently progressive reforms that blur spatial boundaries between state and society, obscure channels of ownership and control, and conceal the identity of state actors. For Cohen, even reforms that seem to mark the retreat of the state and its appetite for control (for example, the development of “community controls” and halfway houses, and the extension of parole) almost invariably augment (as opposed to displace) existing mechanisms of punishment and ultimately expand state prerogatives. A similar logic inheres in the relationship of the contemporary private prison to sovereignty.
The sovereignty-extending character of the private prison is obvious. The prison, any prison, is an extreme representation of the sovereign prerogative of the state. The private prison is not only a prison, it is (or at least it claims to be) a kind of perfect prison: a more efficient and more effective version of the institution and thus a more efficient and effective articulation of state control. Consistent with this aspect, the rise of the private prison has in no way slowed the rate of incarceration or reduced the scope of the criminal justice system—quite the contrary. When seen in this light, the private prison immediately appears as a development that contradicts the most fundamental ideal behind the rule of law: that of absolutely minimizing the coercive prerogative of the state.

Though troubling, by no means does this straightforward conflict with the rule of law mark the extent of the private prison’s problematic character. Perhaps more insidiously, the private prison’s characteristic interpenetration of public and private translates into a persistent confusion regarding the legal rules that apply to private prisons. As with convict leasing, clear examples of this kind arise around the question of liability. Section 1983 actions are the primary vehicle for the vindication of prisoners’ rights, in particular those concerning conditions of confinement. McKnight resolved negatively the then-unanswered question of whether private prison guards could avail themselves of the “qualified immunity” privilege generally available to state actors who reasonably believe their conduct to be lawful. To some extent, McKnight has clarified the issues and generally increased the likelihood that private prison operators and their agents will be subject to viable inmate lawsuits. But key questions remain unclear, for example, whether such causes of action are equally viable in the federal context (i.e., as Bivens actions). Also unclear is the extent to which the so-called “good faith” defense is available to private prisons in § 1983 cases. Although it is rather well-settled that private prisons constitute state action sufficient to form a basis for such causes of action, it is not at all clear whether any specific deed committed by a private prison or its agents constitutes state action. Although this problem of defining the limits of state action is endemic to civil rights jurisprudence, it seems aggravated in this case by the proliferation of non-state functions and actors in the private prison.

Just as problematic and unclear in the wake of McKnight is the scope of government liability. There is good reason to anticipate that under present law the private prison has the effect of insulating the state from liability and thus legal accountability. This follows not least from the fact that § 1983 generally does not trigger the application of the principle of respondeat superior. In the normal public prison context this situation is problematic enough for its diffusion of liability. In the private prison context, this difficulty is magnified, as the state is one step further removed from exposure to liability (at least to the extent that agents of the state are not “deliberately indifferent” within the meaning of these claims). Notwithstanding McKnight’s clarifying functions, then, the state still seems able to reduce its level of legal responsibility to inmates when it incarcerates them in private prisons. Similarly, it seems probable that the use of private prisons generally limits the litigation expenses of states as well.

It also is likely that the juridical structure of the private prison attenuates and ultimately insulates the state from accountability of a more symbolic, political kind. Private prisons tend to distance public officials from responsibility for the way private prisons are run.
The most obvious evidence of this is that, when private prisons are the subjects of scandal, corruption, and the like, journalists and regulators focus first and most forcefully on the private character of the institution, and only later, if at all, on more general public policy dimensions of criminal incarceration. In similar fashion, the private prison converts the problems of prisons—which are endemic and substantial in every case—into management questions and questions of relative performance, efficiency, contract interpretation, and so forth. Several critics of the private prison have articulated this issue in terms of problems of misaligned “symbolism” and of “intervening” implications of the private prison for the way the public understands the origins and functions of criminal justice policy.

As if this situation did not sufficiently insulate the sovereign from its deeds, there also are complex jurisdictional problems with contemporary private prisons. In particular, private prisons frequently are established within jurisdictions different from the contracting state (i.e., they house out-of-state inmates) or established under contracting regimes that involve intervening governmental entities between the contractor itself and the incarcerating government. Apparently, in Tennessee alone, CCA houses inmates from Washington, D.C., Hawaii, Montana, and Wisconsin. At present, Wisconsin holds the lead with more than 4,000 of its inmates incarcerated in other states. In Louisiana, a chronically abusive juvenile detention center, housing state offenders, operated under a contracting scheme that, by inserting a municipality, left no direct contractual privity between the state and the facility.

In such situations, there remains a great deal of legal uncertainty, which seems to have benefited the contractors more than the states or the wayward inmates. In one notable case, two violent sex offenders from Oregon, housed at a CCA “immigrant detention center” outside of Houston, escaped from the facility. This caused the firm, which initially had declared that it had no obligation to notify the local authorities of the institution’s presence, to proclaim that it was not their function to capture them. When the local authorities captured the inmates, at their own expense, it turned out that because they escaped from a private facility, the escapees could not be prosecuted under Texas law. In a manner slightly reminiscent of state attempts to recover stipulated fines for escapes under convict leasing, Texas since has embarked on a campaign to recover costs of recapture from private prison operators.

The merger of the public and private in the private prison inevitably confuses, as well, the issue of access to private prisons—to prison grounds, records, and so forth. The courts and legislatures have long struggled with access issues in the prison context and have long had to balance the advantages of and legal claims to open access against penological (usually security related) concerns. Foucault describes how all prisons refine the punitive authority of the state by cloaking the mechanisms of punishment in a veil of secrecy. By its very nature, the private prison renders the prison all the more insular and the legal questions surrounding access vastly more complicated, for it adds to the mix the proprietary rights of prison contractors. The issue has not lent, and perhaps cannot lend, itself to any consistent resolution and should be understood as one of several ways the private prison exacerbates the irrationalities of the modern prison.
Yet another area of legal complication centers on the due process rights of inmates incarcerated in private prisons. The law in this area clearly favors the promulgation of consistent, predictable rules and procedures—of the kind, as mentioned earlier, that comport with one of the main normative aspirations of the rule of law and that, as a more substantive matter, ensure that states retain ultimate control over decisions affecting basic terms and conditions of incarceration. The existence of even seemingly mundane rules and procedures are particularly important to the extent that they govern internal affairs of prisons—for example, disciplinary proceedings—that can have substantial implications for length and conditions of incarceration. With private prisons, it seems inevitable to some commentators (not to mention being consistent with the reality of the situation) that the private institutions themselves will, notwithstanding the law, be able to retain substantial authority over these matters, particularly the administration (as opposed to the making) of rules and procedures. To the extent as well that states are unlikely to enact effective safeguards, the possibility remains that the private prisons may be able cynically to sustain their occupancy rates, and therefore their revenues (as most contracts are per inmate/per diem), by manipulating inmates’ terms of incarceration.

Like convict leasing, too, the confused juridical structure of the contemporary private prison is intrinsically connected to endemic corruption. With the private prison, the relevant public and private parties frequently seem to wear the same hat, or live under the same roof, as it were. When, in 1985, CCA attempted to contract with Tennessee for a ninety-nine year management contract covering all the state’s facilities, it turned out that the governor’s wife and the Speaker of the State House owned stock in the firm. Indeed, CCA was founded by politically connected figures: its principal founder was a former chairman of Tennessee’s Republican Party. Similarly, Louisiana’s chronically troubled Tallulah Correctional Center for Youth originally was owned by a group of partners intimately connected to former Governor Edwin Edwards, himself a perennial subject of corruption investigations. Somewhat more direct influence over the political process on the part of private prison operatives also is evident. In Arkansas, the founder of a private prison company recently was sentenced to prison for attempting to bribe a correctional official. In Georgia, CCA invested over $130 million in building 4,500 beds before it had any contracts with the state to house any inmates. Instead, the company had “an understanding” with Georgia’s correctional officials and with local politicians—notwithstanding the state’s competitive bid laws. In that case, Georgia’s Commissioner of Corrections illegally communicated with CCA’s lobbyist on the project throughout the bidding process. In Ohio, CCA recently managed, through a “lobbying blitz,” to defeat legislative attempts to regulate the state’s private prisons. CCA’s lobbyists include, on a national level, J. Michael Quinlan, former Bureau of Prisons Chief under President George Bush Sr., and in Tennessee (again) the wife of the Speaker of the State House. In California, CCA, Wackenhut, and Cornell Corrections recently were reported to have retained some of the state’s “most powerful lobbyists” to expand their market.

While there is not yet any credible evidence of contractors’ attempts, as was the case with convict leasing, to manipulate the criminal law to bolster their business prospects, a finance officer at CCA apparently described the 1994 Federal Crime Bill, with its tougher penalties and grants for prison construction, as something “very favorable to us.” Similarly, a recent conference of private prison contractors featured a keynote address
(by Charles Thomas, whose scholarly contributions are cited in this Article) entitled, “The market remains quite positive.” Suffice to say, there is certainly structural potential for this type of conduct—but given current rates of growth in incarceration, such lobbying is for the moment quite unnecessary anyway.

As was also the case with convict leasing, many private prisons appear unable to insulate decisions about the quality of penological functions from financial considerations. Critics of private prisons continually identify horrendous examples of avaricious, sometimes downright mercenary conduct by prison operators—including underpayment and undertraining of guards and other employees, overcrowding, improper classification of inmates, and patently inadequate security structures. When, in 1994, Human Rights Watch investigated the Tallulah juvenile facility described above, it found not only questionable physical structures and inadequate services, but also that offenders were short of food and provided with grossly inadequate clothing. Company officials there successfully resisted for several years attempts to amend these and other abominable conditions, prompting four temporary takeovers by the state (the last one permanent) as well as extensive litigation. When recently denied a unilateral demand for increased per diem compensation, the facility’s owners (who had just obtained a lucrative and discretionary buy out in the face of a final take-over) simply cut back again on the provision of basic necessities. A newer private juvenile facility in Louisiana, operated by Wackenhut, was also recently made the object of a Justice Department lawsuit alleging, among other outrages, inadequate health care and education, shortages of food, shoes, and bedding.

Indeed, the intrusion of profit motives into management decisions is a pervasive problem with private prisons. The most salient expression of this is that private prison officials inevitably find themselves having to balance separate, often competing interests and sort out competing loyalties. Much of the supposed competitive advantage of private prisons derives from their ability to sidestep the civil service wages required with public prison guards. This dynamic encourages not only the employment of under-trained and disinterested employees but aggregate reductions in staffing—practices which in turn account in part for elevated levels of abuse, inmate-on-inmate violence, and so forth. At each of the private juveniles facilities just mentioned, the Justice Department cited inadequate training, retention, and compensation as contributing causes of abuse. . . .

My point in recounting these dysfunctions is not to rehash the claims of more empirically and practically minded critics of private prisons. Instead, I wish to emphasize that such dysfunctions are neither accidental nor episodic, but instead are intimately related to the absence of rule of law norms. Put another way, these dysfunctions are the predictable companions of a system premised on legal confusion, on divided obligation and interests, on the stealthy extension of the state—premised in short on the thoroughgoing abrogation of rule of law norms and their sovereignty-restraining functions.

C. The Limits of Privatization Versus the Limits of the Prison

To a certain degree, this Article’s critique might seem more an indictment of privatization as such. Indeed, I am admittedly skeptical about most instances of privatization. But
because my arguments against the privatization of prisons focus on the sovereignty-restraining ambition of the rule of law and on the perversion of this ambition by the diffusion and extension of sovereignty, my claims in this Article are primarily applicable to privatization where two factors are present: (1) where the institution in question discharges extreme—that is, especially coercive or violent—sovereign functions; and (2) where the privatized institution retains an especially close connection to the state, the state’s interests and its functions. Such characteristics are, as we have seen, especially evident with the private prison.

To some extent, of course, virtually all institutions that can be privatized are coercive and entail the exercise of sovereign-like functions and functions which otherwise could be performed by the state. To some extent, one also might argue, virtually all privatizations remain connected to the state. Such statements parallel the truth, so well exposed by critical legal scholars, that the public-private distinction never is complete in any given direction anyway. Accordingly, it is quite impossible to rigidly circumscribe the limits of this Article’s critique. Nevertheless, significant quantitative differences prevail between the levels of coercion, of sovereignty, and of state presence evident with prisons versus, for example, schools and utilities. In other words, the prison is unquestionably extraordinary in its level of coerciveness and in its extreme representation of sovereignty. Therefore, whether or not other types of privatization are problematic (and again I think that they usually are), there are specific reasons why the private prison is especially problematic from a rule of law standpoint.

As is the case in all contexts where the fate of liberal legal norms are at stake, the benefits accorded by adhering to the rule of law where prisons are concerned are relative and contingent. The public prison remains intensely problematic and in many ways inherently irrational. Rather than offering some romanticized defense of the public prison, I conceive of this critique as a way of suggesting that there are inherent, structural reasons to suppose that private prisons will always, on the whole, remain more dysfunctional and indeed more socially malignant than public prisons. But perhaps more fundamental from a rule of law standpoint is the idea that at least the public prison is transparently problematic and irrational, and at least it requires the state to face directly the political, legal, and fiscal costs of pursuing a criminal justice policy that has brought about almost exponential increases in the rate and the aggregate number of people incarcerated. Indeed, in a society that claims a basis in rule of law norms, it is probably always a good thing for the state to wage its own wars against its citizens and to do so in an obvious and maximally costly way.

VI. RULE OF LAW, THE PRIVATE PRISON, AND THE INEVITABLE SPECTER OF ILLEGITIMACY

In the end, it may be that the problem of privatization is no more than a reflection of the irrationality of contemporary society and a reflection, too, of what Roger Cotterrell, following Neumann, has called the “largely unfavorable conditions in which the idea of Rule of Law seems to exist today.” Such seems certainly true of the private prison, which is at root a sub-species of an institution that is, in the final analysis, fundamentally irrational. Put another way, we might think of the private prison and convict leasing as together demonstrating the inability of the legal structure of the liberal state to retain its
coherence in the context of fundamentally illiberal practices in punishment. But as always, there are real choices between unattainable ideals and the brutal reality of the status quo. From this standpoint, the private prison emerges as something worthy of great concern precisely because of the inherent irrationality of prison as such.

Finally, I do not deny that aggressive courts, competent legislatures, and zealous reformers theoretically could resolve all the diverse problems that plague private prisons: the uncertainty about liability and the line between the state and the contractor, the problems of accountability and public perception, the jurisdictional problems, and so forth. But if convict leasing suggests anything about private prisons it is first that juridical structure is relevant to the prospects of reform, and second that the possibility of reform must not be confused with its probability. Of course, for those who approach this issue in a more principled, or at least more skeptical way, this is all beside the point anyway, since it is clear that reforming and clarifying the legal and political character of an institution premised on the merger of the public and private only can be accomplished by legalizing the interpenetration of public and private and by affirming the normative dysfunctions that come with the private prison. For, while the institutions’ various dysfunctions—corruption, abuse, confusion about liability—seem quite logically related to a lack of state regulation and control, increasing the involvement of the state in the operation of private prisons, short of abolishing private prisons as such, can only have the effect of more deeply entrenching the juridical dynamics—the interpenetration of public and private and the diffusion and extension of sovereignty—that underlay the private prison’s problematic character in the first place. From a rule of law standpoint, the private prison seems a hopelessly problematic institution.

§ 9.3. Foreign Affairs Privatization


I. INTRODUCTION

. . . The end of the Cold War produced dramatic changes in the complexion of international relations and in the evolution of international law. The demise of the Soviet Union and the fall of communist regimes in Eastern Europe unleashed long-suppressed ethnic rivalries and historic conflicts resulting in low-scale wars, most notably evidenced in the Balkan conflict. In other regions of the world, low-intensity, internal conflicts now dominate the geopolitical scene but receive relatively little attention or influence from foreign governments. The great states that once dominated the international landscape and frequently intervened in local and regional conflicts are now absent. Instead, a geopolitical power vacuum has replaced the bipolar international scene.
It is in response to this power vacuum that sophisticated military professionals have created the modern, private, international security company (SC). SCs provide a wide spectrum of military and security services usually reserved for official militaries. The services they offer range from creating military doctrine or manuals for established national militaries to engaging in actual combat in ongoing conflicts. These private companies have formed in militarily advanced countries, relying on the excess of military expertise and extensive networks of retired military officers to provide their services. Consequently, SCs have strong personal and professional links to the governments and militaries of their respective home states. They often work for their home-state governments and contract with foreign states, usually training national militaries. Some of these companies form parts of larger corporations with extensive economic interests. Most SCs have enjoyed enormous success and growth in this decade.

Two such SCs have received widespread attention in the world press because of their involvement in volatile military and political situations and their government contracts. Executive Outcomes (EO), a South Africa-based company composed of ex-commandos who served in South Africa’s apartheid-era security forces, gained notoriety with its direct involvement and success in the Angolan and Sierra Leonese civil wars and its canceled contract with the government of Papua New Guinea. Military Professional Resources, Inc. (MPRI), a U.S.-based company founded and operated by former high-level U.S. military officers, contracted with both the Croatian Federation and the Bosnian Federation, among other parties, to train and professionalize their respective militaries.

Most commentators on these companies have characterized SCs as sophisticated mercenaries. Indeed, SCs have met with the same sorts of criticism meted to mercenaries in the past. EO’s direct military involvement in African civil wars, for example, has met with severe criticism due to the company’s connection to the apartheid-era security forces that intervened in African conflicts and covertly attacked established regimes. In addition, EO’s direct corporate ties to multinational energy and mining companies have frightened those concerned with the corporate group’s overwhelming resources, ability to control valuable territories, and influence over contracting governments.

MPRI has also come under international scrutiny. The media credited the company with resurrecting the Croatian military as it retook crucial territory, such as western Slavonia and the Krajina region, previously held by the Serbs. The same observers have criticized the company’s current contract to “equip and train” Bosnian Federation forces as an arrangement that inflames tensions in the region and puts NATO soldiers at risk.

Concern about these SCs, like concern about mercenaries, pirates, and terrorists, stems from the inherent violence of their profession combined with a lack of control over and accountability for their actions. Since these are private companies, countries which recommend or export them arguably can disavow any connection to SCs’ activities. Potentially, this allows exporting governments to use SCs as political pawns to affect the internal affairs of a country or region while retaining their official neutrality in such conflicts. Such “SC exportation” seems like a convenient way for exporting countries to intervene with impunity in foreign wars.
Meanwhile, countries importing SCs gain quick access, with seemingly no political strings attached, to needed military expertise during a crisis. The contracting country can use and dispose of these services readily without concern for the company’s political ambitions or for political favors which may need to be repaid. This “clean hands” approach to foreign policy appears dangerous to those who see transparent nation-state accountability as essential to controlling human rights violations and the type and quality of military activity throughout the world.

In terms of SCs’ motivations, there is concern that they act only in their pecuniary interests and change allegiances easily and often. Given their high-tech expertise and hardware, their sophisticated communication and tactical skills, and their ability to coordinate military operations, these companies could be hired by insurgents or foreign governments to help destabilize a recognized regime or to suppress a legitimate national liberation movement (groups fighting racist or colonial regimes). In the case of SCs that have links to multinational corporations with significant interests within contracting countries, the concern is that SCs will act solely for the benefit of those corporations by defending only their property, resulting in the creation of semi-sovereign entities to which the contracting government is beholden. Thus far, however, SCs have avoided such criticism by contracting only with legitimate, internationally recognized regimes. In addition, SCs have restricted their activities to training national militaries as opposed to serving as infantry in battle. In certain regions, they have even conducted humanitarian work alongside foreign forces.

The existence of SCs calls into question their legality. International law on mercenaries, as captured in U.N. resolutions and state action, has failed to crystallize. To the extent that such law can be identified, it provides little guidance regarding the services SCs provide, whether training or actual combat. First, there is the question of how to define a “mercenary.” Recent codification has tried to define a mercenary, in contradistinction to an international volunteer, by his motivations. In general, a mercenary is defined as a soldier-for-hire, primarily motivated by pecuniary interests, who has no national or territorial stake in a conflict and is paid a salary above the average for others of his rank. This definition has faced criticism from most of the members of the United Nations because of the ambiguity of “motivation.” To many countries there remain subtle, if not artificial, distinctions between a mercenary and other types of combatants, such as “freedom fighters” motivated by ideology or foreigners enlisted in a national army, like the Gurkhas in the British Army or members of the French Foreign Legion.

International attention focused initially on the problem of mercenaries during the 1960s post-colonial era in Africa. Shortly after the Congo gained independence, the United Nations deployed a “peace-keeping” force, United Nations Operation in the Congo (ONUC), to suppress the Katangan secession, which was being aided by European mercenaries. Since that period, mercenaries have been involved in numerous tumultuous episodes in Africa in which they have tried to depose established regimes or otherwise fuel insurrection. These incidents have caused concern that colonial powers or those loyal to them were utilizing mercenaries to regain power in particular countries or to destabilize regimes that were seen by European countries as uncooperative.
As a result, the Organization of African Unity (OAU) galvanized a movement opposing the use of mercenaries, ultimately resulting in U.N. Security Council resolutions banning the recruitment, use, and training of mercenaries for the purpose of destabilizing nascent regimes or supporting national liberation movements. Because the resolutions ban only these particular uses of mercenaries, it is unclear whether the use of mercenaries to protect legitimate governments or “recognized” national liberation movements is illegal under international norms. Although the law against mercenaries seems absolute regardless of their employers or the purpose for which they are employed, the development of this body of law in the wake of African independence from colonial rule calls into question whether certain types of mercenaries may be legally acceptable depending on the legitimacy of the employer and the nature of the service provided.

Furthermore, it is unclear whether a total ban on the use of mercenaries would conflict directly with a country’s right to self-defense, codified in Article 51 of the U.N. Charter, when involved in an international conflict or when engaging an intrastate force that is being aided by a foreign power. Arguably, such a moratorium might also infringe on the nation-state’s inherent right under Article 2(7) to control its own population in situations where a resistance movement does not qualify as a “movement of national liberation.” Thus, there are inherent contradictions between a total ban on the use of mercenaries and the sovereignty rights of nation-states.

Finally, state practice, which in part determines the development of customary international law, does not give rise to an international norm absolutely banning the use of mercenaries. States have been lax in promulgating and enforcing municipal laws that restrict their citizens’ ability to serve as mercenaries. States also continue to hire mercenaries or contract foreigners to achieve their political and military needs.

The thesis of this Article is that the customary international law banning the use of mercenaries should not apply to SCs that are hired by legitimate governments or by internationally recognized movements of national liberation for either training or combat support. The international community has developed its laws on the use of mercenaries so as to avoid their involvement in attacking sovereign, legitimate states, suppressing movements of national liberation, or disrupting a population’s right to self-determination. SCs cannot be considered “mercenaries” because their activities have not challenged the sovereignty of states or the right of populations to self-determination. Instead, SCs have restricted their contracts solely to work for legitimate regimes or organizations. The laws banning mercenaries do not apply to these companies when they are employed in such a capacity.

However, the geopolitical market may change so that the only available contracts for SCs are with disreputable governments involved in controversial conflicts or with insurgencies. In light of this possibility, national governments should design specific regulations for SCs in order to temper their potential disruptive effects. Licensing and registration procedures, in addition to periodic reporting requirements, will enforce a degree of transparency and link SCs’ activities to their employing states. When governments exert control over SCs’ profitability or ability to operate efficiently, SCs will become attuned to the needs of the state and will align their operations with the
state’s interests. For the international community, this means that there will be a greater degree of accountability for SCs’ actions; consequently, SCs will not pose the same threat as traditional mercenaries.

Part II of this Article will survey the history of mercenaries and examine where modern SCs fit into this evolution. It will show that modern SCs resemble Italian or free companies which surged during the Italian Renaissance under similar sociopolitical situations. Part III will examine the operations of SCs in depth. In particular, it will focus on EO and MPRI to give the reader a sense of how these companies function, what services they provide, where they have been successful, and their relationships with their respective governments. Part IV will analyze the international law regarding mercenaries. It will focus on the historical context of these laws and discuss the problematic nature of defining mercenaries by their motivation. Part V will confront the normative questions regarding the growth and popularity of SCs and whether regulation should be the responsibility of the international community or of individual states. It will provide some recommendations for the treatment of these companies. Part VI will examine the reaction to and regulation of SCs by their respective exporting countries. In particular, it will dissect the licensing procedure used by the United States to monitor companies providing defense services.

In sum, this Article will analyze the implications of the emergence of private military force in the form of SCs. In a world suffering from increased low-intensity warfare that threatens the nation-state system, the United Nations and the major powers have reacted slowly and ineffectively to the security needs of smaller states. This new world disorder has given birth to SCs, which act as surrogates for state power.

II. HISTORY OF MERCENARISM

To understand the emerging forms of international security companies and place them in proper historical context, it is necessary to survey the evolution of mercenaries. SCs represent a reconstituted form of organized corporate mercenarism that is responding to the need for advanced military expertise in escalating internal conflicts. SCs also present a new means of disguised efforts by their home states to influence conflicts in which the home states are technically neutral. In this sense, the emergence of SCs is not a revolutionary development in military and geopolitical strategy but a permutation of past forms of mercenarism adapted to the demands of the post-Cold War world.

In addition, the historical context of the development and modern existence of mercenary troops is essential to an understanding of the modern state response to mercenaries and the resolutions and conventions adopted by many countries to regulate or outlaw their activity. To some degree, this context will inform and dictate how states react to the new SCs domestically and as members of the international community.

As long as humanity has waged war, there have been mercenaries; only in the twentieth century has the mercenary been vilified and outlawed. The forms of mercenarism have changed through the centuries, but the basic character of the soldier-for-hire has survived. Even in the twentieth century, when nation-states have monopolized the domestic and
international uses of force by establishing conscription and the citizen-army, mercenaries have thrived. In general, mercenaries have appeared where there has been a breakdown of internal order in a country. For states or entities caught in tumultuous civil disorder or war, mercenaries have been a source of instant military force and expertise. Countries that produce mercenaries can affect the internal balance or regional composition of an area of the world while claiming neutrality.

The rulers and challengers of ancient empires and regimes used mercenaries. In Anabasis, Xenophon recorded the first account of mercenarism—the failed use of 10,000 mercenaries in 401 B.C. by Cyrus, a pretender to the Persian throne. During the fourth century B.C., Greek city-states introduced specialized noncitizen warriors into intra-Greek battles. The Macedonian successors to Alexander the Great utilized highly trained mercenary armies to wage war. The Roman Empire relied on native populations like the Teutonic tribesmen to expand frontiers and control territory. The rulers of Byzantium and Carthage also depended on the military expertise of foreigners to fight their battles.

Kings and lords in the Middle Ages also hired mercenaries to fight for them. Since knights were only required to serve forty days a year and were not obligated to serve abroad, rulers were forced to hire soldiers when they wished to launch offensive military expeditions. In the twelfth century, the English king introduced the system of scutage which allowed individuals to buy their way out of military service, thereby giving the king the necessary funds to buy professional military forces; English kings thereafter contracted with noblemen to provide the service of their retinues beyond the customary feudal obligation. Kings also relied on private or royal subcontractors to raise and supply armies for them.

This continued reliance on hired soldiers spawned the creation of mercenary “free companies,” which were bands of fighting men who offered their military skills jointly. The Byzantine Emperor hired the first such company, the Grand Catalan Company, to fight the Turks. The company’s 6500 soldiers were recruited from the Aragon mountain region and were led by a German, Roger von Blum. In 1311 the Grand Catalan Company betrayed the Duke of Athens and established a duchy in Athens, where they survived for sixty-three years. Other free companies entered the mercenary market employing soldiers from all regions of Europe.

In Italy, free companies phased out as the northern city-states prospered and began contracting with freelance commanders, or condottieri, to supply specific numbers of troops for particular military services. For the Milanese dukes, Venetian doges, the Queen of Naples, Florentine financiers, and the Pope, it became safer to hire professional outsiders, under business-like contracts called condotta, than to employ potential rivals from their respective domains. In addition, hiring well-equipped and trained professional fighters did not disrupt or distract the productive economy by forcing normal citizens into military service.

When waging war, the condottieri concentrated on taking prisoners, since their preoccupation was with raising ransom money. Generally, since they were ideologically and politically detached from their battles, the hired soldiers were not interested in killing
per se; instead, they conducted themselves within the accepted professional strictures of warfare. As technological innovations like light firearms appeared and larger scale armies engaged in total warfare, the condottieri slowly gave way to national armies composed of citizens and independent mercenaries.

In the seventeenth and eighteenth centuries, European states began to hire soldiers and sailors from all regions to serve in their respective armed forces. The Swiss were a major supplier of troops and officers, especially to France. Swiss cantons (political subdivisions) like Uri, Unterwald, and Schwyz developed a system in which foreign rulers would pay the canton directly for the services of cantonal armies. German princes similarly leased their troops to foreign powers. The Dutch both employed and provided mercenary forces and created the Scots Brigade, a legion completely composed of foreigners. Britain also hired mercenaries, especially from Germany, to fight its battles in Scotland and on the continent. During the American Revolution, the British tried to hire Russian troops and the Scots Brigade, but their requests were rejected. The Crown instead employed German troops including Hessians, the most lethal and well-trained of the German mercenary forces. By the middle of the eighteenth century, there were three ways to obtain the services of mercenaries: direct enlistment, purchase or lease of regular army units, and the subsidization of another state’s army.

In the nineteenth century, with the consolidation of central authority in most European states and the establishment of the notion of nation-state sovereignty, the recruitment of foreigners for duty in national armies declined. States began to pass neutrality laws which generally prohibited their citizens’ enlistment in foreign armies, and citizen-armies became the norm. Nation-states therefore began to monopolize the authority to deploy force abroad and to accept responsibility for violence emanating from their own jurisdiction. Because of this, mercenaries were no longer seen as merely independent agents selling their services in the market but as parties “lending [their] skills in the application of coercive force to some political cause.” But by this period, mercenaries also had proven themselves unreliable, as some refused to fight and others deserted.

The last instance when a European state raised an army of foreigners to fight in a European war was in 1854, when Britain hired 16,500 Germans, Italians, and Swiss mercenaries to fight in the Crimean War. However, a tacit code against the use of mercenaries on the European continent did not stop European states from employing foreigners to wage their extra-continental battles. Since the defeat of Nepal and the Treaty of Segauli in 1816, the British have recruited Gurkhas into infantry regiments and used them in their colonial wars. In 1831, King Louis-Philippe of France enacted a law which established the Legion Étrangère, the Foreign Legion, composed of foreign volunteers. France has used these well-trained regiments of foreigners in many of her colonial battles, from Indochina to Africa.

In the twentieth century, aside from the enlistment of foreigners directly into nation-states’ standing armies and the seconding of officers into the armed forces of ex-colonial regimes, there are two other ways in which foreign troops are employed in the twentieth century. The first is the hiring of another state’s troops on a per capita basis, in which states negotiate the exchange and use of a specific number of national troops for a fixed
price. For example, the United States hired Australian, New Zealand, Thai, Filipino, South Korean, and Chinese troops to fight in Vietnam. The second and more popular form is when states contract directly with individual foreigners for their services in a particular conflict, with or without the consent of the exporting state. These independent mercenaries, commonly referred to as “soldiers of fortune,” “wild geese,” or les affreux (“the dreaded ones”), have been the focus of much attention since the 1960s, when they appeared in post-colonial Africa.

When they first appeared in Africa, these independent mercenaries, hired outside the constraints of the twentieth century nation-state system and seemingly motivated solely by pecuniary interests, were seen as a shocking anachronism. In addition, they presented a threat to the fledgling independence of the African state. It was only twelve days after gaining independence from Belgium that secessionist leader Moise Tshombe declared independence for the southern mining province of Katanga in the Congo and, supported by Belgium, hired white mercenaries to help him establish his regime and fight the Congo’s main force, the Armée Nationale Congolaise (ANC). Those opposed to Tshombe’s actions saw the Katangan secession as an attempt by Western capitalists to retain control of a mineral-rich region. In its first attempt at peace enforcement, the United Nations responded by sending U.N. troops to fight the mercenaries. The United Nations crushed the secession in 1963 and departed in 1964. The mercenaries, however, had fought convincingly and forced the international community to take notice of the resurrection of the independent mercenary.

In Africa, the mercenary threat to weak nation-states struggling to establish their independence and legitimacy became very clear as independent mercenaries continued to influence intrastate conflicts. In the Congo, Tshombe, whom President Christophe Gbenye appointed prime minister, hired mercenaries to crush the Simba revolt in 1964. Later, in November 1967, General Mobutu Sese Seko ended a mercenary revolt led by Bob Denard, an ex-officer in the French marines, and Jack Schramme, a Belgian, who were commanders of separate mercenary commando units in the Zairean armed forces.

In 1967, mercenaries joined both sides in the Nigerian Civil War. A Pretoria-based company, Mercenaire International, supplied mercenaries to the secessionist region of Biafra, which ultimately lost in 1970 after three years of intense fighting. Mercenaries entered the conflict in the southern Sudan in 1969. Portuguese-led mercenaries attacked the socialist government of Guinea in November 1970. Also during that decade, mercenaries enlisted in the military forces of Holden Roberto’s right-wing National Front for the Liberation of Angola (FNLA) and Dr. Jonas Savimbi’s Union for the Total Independence of Angola (UNITA), while Cuban troops were sent to support the socialist government of President Agostinho Neto. Then the capture and trial of thirteen mercenaries galvanized the African movement against mercenarism yet again and forced supplier nations to take notice of the problem of recruitment.

There were several other dramatic coups and coup attempts involving mercenaries in the 1970s and 1980s, including the consecutive overthrows of the government of the Comoro Islands by Bob Denard in 1975 and 1978 and the attempted overthrow of the regime in the Seychelles by Mike Hoare in 1981. In 1976, opponents of President Mathieu
Kerekou’s regime in Benin (Dahomey) hired Denard to launch an attack and establish a new regime. Denard’s Force Omega, composed of ninety mercenaries, failed to assassinate the president, were met by stiff public resistance, and ultimately were forced to abandon their million-dollar operation. The Organization of African Unity responded to all of these incidents with declarations of condemnation and called on all nations to outlaw the recruitment and use of mercenaries.

In most cases where mercenaries were involved, there seemed to be vital economic interests at stake, usually mining and oil interests. For African countries, this pattern further proved that mercenaries were simply tools of colonial powers who had economic interests vested in their continent. With pressure from African countries, the U.N. Security Council passed several resolutions in the 1960s condemning the recruitment, use, or support of mercenary troops against newly independent countries or to suppress national liberation movements.

Even though there was general distaste for the reliance on foreigners, countries in Africa requested auxiliary troops from other nations to support their fledgling regimes. There developed a clear distinction between foreign support of legitimate African regimes and individualized mercenary attempts to wreak havoc in the region. For example, Great Britain sent its troops to restore order in Kenya, Uganda, and Tanzania in 1964, and France deployed forces to Gabon at the request of the ruling regime. In general, the reaction against the use of mercenaries, seen as unaccountable, self-interested henchmen of colonial states, has not been the same for nationally organized armed forces sent by foreign states to fight for an existing regime.

Since military expertise and demand for military manpower are often at a premium, the use of foreigners in national armies or in internal conflicts has persisted. In the Middle East, Pakistani officers and soldiers serve in several countries’ armies, including Libya’s Islamic Legion. Some former colonial countries continue to second their officers to the armed forces of their former colonies. In other cases, rulers still find it convenient to hire foreigners to act as guards so as to avoid arming their respective nationals and to prevent a challenge to their power.

In the 1980s and 1990s, aside from launching coup d’etats against small African regimes, mercenaries have provided military force to insurgencies and entities involved in civil conflicts. Mercenaries from Israel, the United States, Great Britain, and France have thrived in regions like Central America and Lebanon, where there is a general deterioration of state control and a demand for fighting power. In Colombia, drug traffickers have hired Israeli and other mercenaries to protect their trade and enforce their de facto rule. British Petroleum has hired Colombian soldiers to guard its facilities against guerrilla attacks. In Haiti, the wealthy have recruited soldiers to form private family forces. Foreign corporations in Liberia create industrial gangs that extract natural resources. The trend of privatizing violence in conjunction with private security and commercial interests continues to grow.

Seemingly, as long as there is war, there will be a need for military expertise. Furthermore, those states from which the mercenaries are supplied have a vested interest
in retaining the option to influence foreign conflicts by allowing mercenaries to sell their services without the possibility of denying responsibility for their actions. It is in this historical and current context that the emergence of the international private security company, in its varied forms, must be understood and analyzed. Without this background in mind, the legal developments regarding mercenary troops and state responses to the recruitment, use, and deployment of mercenaries cannot be understood completely.

III. The Private International Security Company

The private international security company (SC) is a recycled form of past mercenary organizations, like the condottieri and the free companies, that contracted their soldiers out to foreign entities. Although in form resembling their antecedents, SCs have developed a modus operandi compatible with the needs and strictures of the post-Cold War, state-based international system.

SCs primarily have responded to the geopolitical demands of smaller states which have been forced to seek alternatives for the containment of internal dissent or the development of military preparedness in the absence of superpower support. At the same time, SCs have restricted their employers either to recognized regimes in low-level conflicts in which the insurgency has not gained popular support or international recognition, or to regimes supported by their home states. Understanding the disaffection of the international community with the use of force by private actors, SCs have curtailed their military engagements and have strictly controlled the actions of their employees in the field. In addition, SCs have recognized that their survival hinges on the positive perceptions and approval of their actions by their home governments and by the international community; therefore, they have sought government approval for their activities and have conferred with officials during ongoing operations.

Consequently, SCs have become a type of state agent—tied to their home state by tacit or licensed approval for their activities and enlisted as contractors for the employing country. Hence, state responsibility for the actions of SCs seems to flow in two directions: responsibility toward the home state which tolerates and “exports” these companies’ services, and responsibility toward the contracting state which enlists and directs the activities of the SC. States on both sides of the contract may attempt to disavow any connection to particular SCs, claiming that the SCs’ activities are private. To the degree that connections between state acquiescence and SC activity may be drawn, the problem of lack of accountability for private actions in the international arena is vitiated.

As noted above, SCs have sprouted in countries with a large pool of military expertise, due to military downsizing, early retirement incentives, and the sheer profitability of such work compared to regular military pay. SCs have developed distinct market niches tied to the interests of their respective home states. They provide a range of services, from development of training manuals to actual combat service. The panoply of services defies classification, but they all involve the export of private military expertise in some fashion. The trend in the international sector is for SCs to provide military training to government troops involved in volatile (or potentially volatile) conflicts. Though this
trend does not involve actual offensive engagement by SCs, this type of contract is controversial because it implicates the injection of private, foreign military expertise into a precarious arena, which could be determinative of that country’s fate.

A. South African Security Companies

The most controversial of the SCs is South Africa-based Executive Outcomes (EO), founded in 1989 as a limited liability company registered in Great Britain and South Africa. The founder, Eeben Barlow, is a charismatic former 32nd Battalion commander and former top official at the Civil Cooperation Bureau. EO offers general security services, military training, infantry troops, and air support, and its recent successes have led some to call EO the “most deadly and efficient force operating in sub-Saharan Africa today [aside from the South African army].” A firm brochure states that EO provides “the most professional training packages currently available to armed forces, covering aspects related to land, air and sea warfare.”

EO assembles teams of military personnel for each contract, garnered from a list of over 2000 former members of the apartheid-era South African Defense Force (SADF) and South African Police. EO selectively recruits former members of the notorious and feared special-forces units like the 32nd Battalion, the Reconnaissance Commandos, the Parachute Brigade, and the paramilitary Koevoet force. These forces were often used by the apartheid regime to covertly destabilize neighboring countries and to thwart opposition internally. Former members of the SADF are attracted to employment with EO because it offers high salaries, medical and life insurance benefits, and the opportunity to use their military training. Though whites predominate in the officer corps of EO, approximately seventy percent of EO’s soldiers are black.

EO proclaims that it is an apolitical force that will work only for legitimate governments and not for rogue regimes like Sudan or patently unpopular regimes like the former Mobutu regime in Zaire. Though EO claims that it merely trains soldiers or engages in defensive pre-emptive strikes, EO has openly engaged in battles and introduced modern weaponry and tactics with devastating effects in two civil wars. In Angola and Sierra Leone, EO’s superior technology, skill, and collective experience proved crucial in forcing the rebel movements in each country to negotiate respective settlements and in restoring social order.

In 1993, the Angolan government, controlled by the Marxist Movimento Popular de Libertaçao de Angola (Popular Movement for the Liberation of Angola, or MPLA), hired EO for $40 million to help its army defeat Jonas Savimbi’s rebel movement, Uniao Naçional para a Independencia Total de Angola (National Union for the Total Independence of Angola, or Unita). After refusing to accept its defeat in the U.N.-sponsored elections in 1992, Unita resumed its battles against the government. These battles were a mere part of Angola’s ongoing civil war which began in 1975 (ironically, the SADF had fought for Savimbi’s Unita and against the MPLA and their Cuban allies.

---

* [Executive Outcomes was dissolved on January 1, 1999 when South Africa introduced the 1998 Regulation of Foreign Military Assistance Act. —Ed.]
during the Angolan civil war). However opportunistic the switch in allegiances appears, EO used the SADF experience with the Unita rebels to its advantage as it trained Angolan troops and coordinated offensives. EO introduced infrared capabilities (allowing night fighting), advanced communications, reconnaissance, fuel air bombs, and increased air power (in the form of Mi-8, Mi-17, and Mi-24 helicopter gunships and MiG-23 and MiG-27 fighters among other aircraft). Though EO claims that its contract merely involved training Angola’s armed forces, the Forcas Armadas de Angolanas (Angolan Armed Forces, or FAA), EO admits that it engaged in defensive strikes alongside FAA troops which resulted in approximately twenty EO casualties. Its involvement helped the FAA regain crucial mining and oil territory, turned the tide against Unita, and forced Savimbi to sign the U.N.-brokered Lusaka Protocol in November 1995.

As part of the agreement, Savimbi demanded that the MPLA expel EO, a testament to EO’s feared presence. The United States and other foreign governments also pressured the MPLA to end its contract with EO. In early 1996, most of EO’s fighters formally returned to South Africa. Even though it was forced to depart the Angolan market, EO sees itself as the company that was finally able to end the Angolan civil war, and the Angolan government continues to support EO.

Following its success in Angola, EO offered its services to the embattled government of Captain Valentine Strasser of Sierra Leone, then fighting the rebels of the Revolutionary United Front (RUF). The RUF had waged a campaign of social terror since 1991 against the people of Sierra Leone and had occupied valuable mining territory. Sierra Leone’s army, the Republic of Sierra Leone Military Forces (RSLMF), comprised of 3000 poorly trained soldiers, was unprepared to repel the insurgency. Captain Strasser immediately recruited 7000 new, untrained soldiers, many of whom were youths. Economic Community of West African States’ Military Monitoring Group (ECOMOG) forces stationed in Sierra Leone, including Nigerian, Ghanaian and Guinean troops, offered support but proved ineffective against the insurgency. In addition, Strasser enlisted the aid of British SCs, including the Gurkha Security Guards, to help train the new recruits, but they also were unable to stop RUF’s advance.

When Strasser’s government hired EO in 1995, the country was in complete chaos (overrun by organized bandits), and it seemed likely that RUF would overtake the capital of Freetown. Though the government did not have the resources to pay EO immediately, it appears that Strasser offered EO mining concessions in return for its services. For Strasser, this form of payment served to tie EO’s pecuniary motivations with its military success, since EO would have to liberate those strategic mining areas, like the diamond fields in the Kono district, in order to receive its pay.

EO sent approximately 120 military trainers to Sierra Leone to tutor the RSLMF and instill discipline in the motley group of government soldiers; as in Angola, it introduced nighttime operations, and it assumed operational control over offensives while using intelligence reports and sheer firepower to surprise and overpower the RUF’s insurgency. EO launched air assaults against RUF bases with devastating effects, and Nigerian Alpha Jets, from its ECOMOG contingent, often provided tactical air support for EO operations. Some journalists reported that EO’s employees machine-gunned civilians from their
helicopters as they pursued rebels. However, EO has tried to restrain its employees and trainees and has gained a reputation among most of Sierra Leone’s population for being efficient, reliable, and humane.

At the same time, EO threatened to withdraw its services if the Sierra Leone government did not honor its commitment to hold elections as promised. As the RUF forces retreated into Guinea and Liberia and EO retook key territories, the Sierra Leone government held parliamentary and presidential elections on February 27, 1996, and March 15, 1996, respectively. The new president, Ahmed Tejan Kabbah, negotiated a cease-fire agreement and engaged in peace talks with RUF. The rebel leadership admitted that EO was ultimately responsible for its defeat, and RUF demanded that EO, as well as the Nigerian forces, be expelled before negotiations continued. In December 1996, the parties to the conflict signed the Abidjan Accord, calling for the demobilization and disarming of RUF and the cessation of all EO’s activities.

EO also played a role in the internal politics of Sierra Leone. Besides tying its continued service during the civil war to the holding of elections, EO supported the Kabbah administration against disgruntled RSLMF officers whose illicit diamond trading was threatened by the new government. EO made it clear to all internal parties that it was contractually bound to the elected, legitimate government of Sierra Leone and declared that it would defend it against any coup attempt. This allegiance led Sierra Leone’s envoy to Washington to state that “[t]he government of Sierra Leone believes EO can do a better job [providing security] than the Sierra Leone army.” By throwing its support behind the newly elected government, EO was not only concerned with fulfilling its contractual obligations, but was laying the groundwork for stability and establishing its legitimacy in the eyes of future clients and the international community.

EO’s humanitarian work in Sierra Leone distinguishes it from the stereotypical mercenary. EO coordinated and led the return of schoolchildren and teachers, who had been trapped in the capital due to the danger of travel, to their homes in the southern and eastern parts of the country. Thousands of other displaced residents have returned home since the elections. EO also organized the integration of hundreds of child-soldiers from both sides into a rehabilitation project. These services, in addition to the perceived incorruptibility of EO’s forces, endeared the South Africans to most of the Sierra Leone public. By the time EO departed in February 1997, it had established itself as the company that restored order and democracy to Sierra Leone.

Since EO’s departure, however, the nascent Sierra Leone democracy has fallen to a military coup, and the country is mired yet again in a state of destruction and banditry. The chaos which followed EO’s departure is a testament to its effectiveness in maintaining order. It is also evidence that the services SCs provide are not a cure for the many ills which plague war-torn countries.

In both Angola and Sierra Leone, EO conducted itself professionally and compiled a respectable human rights record, especially relative to other African armies. Its successes and popularity in Angola and Sierra Leone have allowed EO to proclaim that it is “trying to aid growth and democracy by bringing stability and foreign investment.” EO attracts
considerable media and international scrutiny by continuously averring it will work only in a defensive/training context for legitimate governments. Though detractors may find it difficult to believe this rhetoric, EO has certainly constrained its employment options by avoiding hypocrisy thus far.

In late February 1997, the government of Papua New Guinea (PNG) entered into a $36 million contract with Sandline International, a subsidiary of EO, to train and equip its soldiers in their nine-year battle against the secessionist Bougainville Revolutionary Army (BRA). The reported one-year contract called for Sandline to both be a “force multiplier” and lead the assault on BRA leaders. PNG’s prime minister, Sir Julius Chan, approached EO because the Papua New Guinea Defense Force (PNGDF) was ill-disciplined, repeatedly engaged in human rights abuses, and was unable to force the BRA to negotiate a settlement.

Australia, PNG’s former colonial master and current military benefactor, protested the Sandline contract fearing a loss of its influence and the introduction of South African “mercenaries” into the civil conflict. In recent years, Australia has vacillated in its foreign policy toward PNG, continuing to provide military and general aid (worth $334 million combined annually) while denying the PNGDF any specialist training or sophisticated weaponry and blocking any equipment transfers from other Western states for their fight against the BRA. Chan’s frustration with Australia’s perceived lack of support and its ability to hinder any other source of aid compelled the prime minister to search for military expertise in the private international market.

Though Chan’s government felt the pressure of international criticism from Western states, its contract with Sandline faced greater opposition from its own military and the general population. The PNGDF commander, Brigadier General Jerry Singirok, refused to work with the hired trainers and demanded the resignation of the prime minister (although General Singirok had helped negotiate the Sandline contract). Chan sacked General Singirok, but the general remained in control of his troops and continued to demand Chan’s resignation. Mass protests and social unrest forced Chan to cancel the contract and send home the EO soldiers who already had arrived in PNG. Chan ordered a judicial review of the Sandline contract, but he also appointed a replacement for General Singirok. After days of rioting on the streets of Port Moresby and continued allegiance on the part of the PNGDF troops to General Singirok, Chan temporarily resigned on March 26, 1997, awaiting the judicial inquiry into the Sandline contract.

The canceled PNG contract may lend insight into the future of the SC market. First, the protests lodged by foreign nations and internal opposition focused on the direct offensive role to be played by EO soldiers, especially the alleged plan to decapitate the BRA by assassinating its leaders. The international community and societies still cohesive enough to protest government decisions may be averse to the direct, aggressive involvement of foreigners in an ongoing civil strife. Second, the internal protest from the PNGDF may indicate that contracts for foreign military assistance may not be appropriate or acceptable in countries which still have a relatively well-structured military with the self-perceived capability of waging an effective war. Third, the international pressure may indicate that SCs provide small states with the independent capability to fight their own
battles, without having to depend on a hegemon like Australia. More than likely, Australia feared a loss of influence in its former colony, and for that reason it adamantly protested the presence of EO. Therefore, international protests of SC involvement should be viewed through a realist’s prism of large states versus small states.

Aside from its contracts in Angola and Sierra Leone and its canceled project in PNG, EO is involved in several African countries in various capacities. EO forms part of a tangled, constantly shifting web of corporate interests organized under the Strategic Resources Group (SRG). These corporations, which include mining, oil, infrastructure, air transport, hospital construction, demining, water purification, computer software, and other businesses, feed off each other in symbiotic fashion. Several reports linked EO with British and Canadian companies and prominent U.K. business figures like Tony Buckingham and Sir David Steel, the former Liberal Party leader.

This veiled relationship with exploitative industries has led many to construe EO (“the Diamond Dogs of War”) as part of a new brand of mercantile company, providing stability for countries in exchange for economic concessions and control of productive territories. Its major contracts to date have been with countries attempting to regain valuable mining or oil installations, leading many to suspect that EO will fight only to gain concessions for its corporate brethren. Other observers fear that the corporate group’s concentration of resources, which allows EO and its affiliates to act as a semi-sovereign force, poses grave dangers to weak African states beholden to capitalist interests. In addition, EO’s ties to local security companies like Lifeguard allow members to remain in a country to provide commercial security for EO interests even after the EO contract has expired—which seems to have happened in Angola and Sierra Leone. Even so, in countries where the control of strategic economic resources is central to the retention of power, the marriage of EO’s military services with corporate commercial concerns makes sense for contracting governments. Many foreign commercial interests scramble for access to valuable resources, so it seems no less egregious for EO and its related companies to take advantage of the concessions offered by the government in return for peace. Furthermore, EO’s contracts are transient, but its commercial interests are long-term. EO has little market incentive to threaten a government with which it is working and has yet to turn on its employers.

EO’s combination of military power, economic resources, and secretive operations is feared by many countries in Africa, which still associate EO with the destabilizing apartheid-era forces of the interventionist SADF. Although EO claims to work only for legitimate governments, rumors continue to persist that EO is engaged in destabilizing activities. Many reports stated that EO helped Mobutu’s forces in Zaire against the advancing rebel force of Laurent Kabila’s Alliance of Democratic Forces for the Liberation of Congo-Zaire (ADFL). EO denied its involvement, and Eeben Barlow claimed that Zaire never approached the company. He averred that EO was constrained by its agreement with Angola never to work with the Angolan “opposition,” which included Mobutu. In addition, Barlow claimed that EO does not enter conflicts where the populace does not support the army (in essence, when the government lacks any semblance of real legitimacy). However, Namibian Defense Minister Phillemon Malima claimed in April 1997 that he had documentary proof that EO was training Namibians in...
Sierra Leone in order to destabilize Namibia. EO denies the statement, and Malima has yet to prove any EO complicity.

All of this exposure has led the South African government to disavow the activities of EO and to propose new legislation to outlaw or severely restrict the activities of South African SCs. The legislation, a proposed amendment to the Defense Act called the Foreign Military Assistance Bill, would expand the definition of “mercenary” to include security services and military consulting and would only allow the export of South African military expertise with the express consent of the South African government. South African SCs have operated without the official consent of the South African government because their activities have not been considered “mercenary” under the provisions of the Defense Act and because their activities were conducted abroad, beyond South African jurisdiction....

Regardless of the proposed legislation, the extreme success of EO and the worldwide respect for the military prowess of ex-SADF fighters and intelligence personnel has led others to consider mimicking EO’s business concept. Since 1992, former SADF officers have planned the creation of an organization to compete with EO. They plan to offer military expertise, obtained from thousands of soldiers made redundant by the SADF, to foreign governments and international organizations operating in war zones. Additionally, three leading officials from South Africa’s National Intelligence Service (NIS) have formed a private intelligence and security service, Panasec Corporate Dynamics (PANASEC). The company hopes to capitalize on its extensive government connections to retain a close working relationship with the South African government. Another company, Strategic Concept, hires ex-soldiers to serve as combat trainers. In general, South African SCs continue to grow by keeping a close relationship with the South African government for business and legal reasons.

B. United States Security Companies

Currently, no U.S. companies provide combat units for ongoing civil battles as does Executive Outcomes. There are, however, several U.S. companies which offer security-based services abroad. Two companies in particular provide military training and organization services to foreign governments.

During the Vietnam War, Vinnell Corp., a subsidiary of Virginia-based BDM International Inc., supported the U.S. military and built military bases, although some Army officers claim that Vinnell also operated clandestine programs such as providing rear security for retreating U.S. forces. Since 1975, Vinnell has trained the Saudi National Guard, which acts as a royal guard for the monarchy, and other branches of the Saudi military in weapons use, tactics, and logistical operations. The Saudi contract came under intense congressional scrutiny as the Senate Armed Services Committee held hearings to assess its policy implications, but Congress ultimately allowed Vinnell to contract with the Saudis. In 1979, the Saudi monarchy relied heavily on Vinnell to help regain possession of the Grand Mosque at Mecca, which opposition forces had occupied. Allegations persist that since the Vietnam War, Vinnell has provided extralegal means of
achieving U.S. security ends in Central America and the Middle East while avoiding the appearance of official U.S. involvement.

The second company, Virginia-based Military Professional Resources, Inc. (MPRI), specializes in military training, evaluations and assessments, war-gaming, doctrine, simulations, research, and analysis. Organized as a closely held corporation in 1987 by top-ranked retired U.S. generals, MPRI’s product is U.S. military expertise. MPRI services governments and commercial employers by assembling strategic teams compiled from its database of over 2000 former U.S. military professionals. It has capitalized on the expertise of its employees and their professional and personal connections to the U.S. government and military to land contracts earning the company over $12 million in annual revenue. The U.S. government has hired MPRI to create and implement military doctrine, training and education, war-gaming, weapons and systems testing, and logistics support development programs.

Aside from these contracts, MPRI has entered the international market by contracting with the U.S. government for missions abroad and with foreign governments directly. In February 1992, the Department of State hired MPRI for its logistics management capabilities to provide humanitarian-relief supplies and equipment to the “Newly Independent States” of the former Soviet Union. As part of its contract, MPRI coordinated and supervised all aspects of the donation and delivery process while working with foreign governments and the United States. In another contract with the State Department, MPRI provided forty-five monitors along the Serbian-Bosnian border as part of an effort to enforce Belgrade’s ban on the flow of arms and fuel to the Bosnian Serbs.

Most interesting, however, have been MPRI’s contracts with foreign governments to provide military and technical training and force-modernization and -expansion programs. For foreign governments, MPRI represents a private channel through which to gain U.S. military expertise in conditions in which conventional U.S. military assistance programs are not appropriate for political or tactical reasons. Though understood to be a private agency, foreign governments also see MPRI as a quasi-official U.S. military body whose services can represent tacit support by the U.S government. MPRI has worked with Sweden, Taiwan, Croatia, and the Bosnian Federation, and has been approached by former Eastern-bloc countries interested in creating U.S.-style military structures that are more compatible with NATO standards. Some observers claim that MPRI was also involved in Angola, but the company flatly denies doing any work in that country. In 1996, MPRI negotiated with the Sri Lankan government about providing training to the newly created Special Boat Squadron, a Sri Lankan Navy special forces unit. According to MPRI, negotiations broke down after Sri Lanka backed away from the deal.

MPRI’s involvement in the Balkans with the approval of the U.S. government has been the most controversial of the company’s foreign contracts because of MPRI’s injection of military expertise into a region still torn by war. In March 1994, Croatian Defense Minister Gojko Susak sent a letter to the U.S. deputy secretary of defense requesting permission to negotiate with MPRI to obtain U.S. training “in military-civilian relations, program and budget” for its military leaders. The United States allowed Croatia to deal
with MPRI as part of the U.S. Croatian Military Cooperation Agreement. On September 27, 1994, MPRI signed a contract with the Croatian government to democratize the Croatian military and reorganize its troop structure. Croatia also employed MPRI to design and implement institutional and organizational management structures for Croatia’s Ministry of Defense.

The U.S. government reviewed and licensed both of these contracts and deemed them not in violation of the existing arms embargo since they did not involve battlefield strategy, tactics, or weapons. However, some observers of the region claim that MPRI’s training, though seemingly innocuous, was pivotal in the resurgence of the Croatian military and the recapture of critical Croatian territory. When Croatia hired MPRI in the fall of 1994, the Serbs occupied thirty percent of Croatian territory. Shortly after MPRI began training Croatian military officials, the Croats experienced unprecedented success in their campaign to regain territory. In May 1995, the Croats achieved their first important victories by retaking areas held by Croatian Serbs southwest of Zagreb. Later, the Croats engaged the Serbs in western Slavonia in a three-day offensive which resulted in the recapture of that region. In its most stunning offensive, the Croatian military launched Operation Storm in August 1995 to recapture the Krajina region, which had comprised twenty percent of Croat land. The Croats overwhelmed the Serbs with a quick, devastating offensive of 100,000 troops which, according to experts, resembled a U.S.-style attack. Croatia and the U.S. government deny that MPRI violated the arms embargo by involving itself with any aspect of the offensive. MPRI unequivocally denies its involvement and downplays its role in any of the Croatian offensives by emphasizing its strict adherence to the training contract and the license provisions under which it must operate. MPRI also points to the fact that it had only fifteen employees working in Croatia and that it had been in Croatia only a few weeks before the offensives began.

Though it does not seem likely that MPRI was actively engaged in preparing the respective Croatian offensives or in teaching offensive tactics, it does seem probable that U.S. leadership techniques and military organization lectures did help the Croats in their advance. Most observers note a dramatic organizational and attitudinal transformation in the Croatian forces after MPRI began its training, which likely helped Croatian military morale and discipline. Combined with increased stockpiles of heavy arms and weapons garnered in violation of the arms embargo, the Croats undoubtedly used the organizational and leadership lessons learned in the MPRI classrooms to their advantage. By November 1995, the Croats had regained all but four percent of their land and had come to occupy twenty percent of Bosnia.

Tacitly, the United States affected the balance of power in the Balkans by allowing MPRI to train Croatian military officers and civilian officials—a role usually reserved for national militaries. Whether the training involved specific advice in regard to offensive campaigns or tactics is irrelevant to this point; MPRI must admit that its training, though characterized as long-term development, was meant to have some positive effect on the fighting efficiency of the Croatian forces or else the company would not have been hired.

In another sense, the perceived U.S. support of the Croatian military, in addition to the Croats’ use of new weapons, may have affected troop morale on both sides of the
battlefield as Croatia pushed to regain its territory. By licensing MPRI’s contract with Croatia at a crucial period in the Balkan conflict, the United States retained its claim of neutrality while aiding its clandestine ally. In addition, NATO planes bombed key Serbian positions in August 1995, thereby forcing the Serbs to the negotiating table at Dayton.

Even more clearly, the United States proposed that a U.S. security company be used to supply arms to the Bosnian Federation and to train its integrated forces as part of the “Equip and Train Program.” Under the provisions of the November 1995 General Framework Agreement for Peace in Bosnia and Herzegovina (Dayton Accords), Bosnian Federation forces were to be equipped and trained so as to equalize their perceived military disadvantage relative to the Serbs. The United States wanted to equip Bosnia to defend itself once NATO troops and IFOR peacekeepers pulled out. The U.S. government also wanted to ensure the integration of Federation forces. Annex 1B of the agreement, addressing “Regional Stabilization,” states the following:

Recognizing the importance of achieving balanced and stable defense force levels at the lowest numbers consistent with their respective security, and understanding that the establishment of a stable military balance based on the lowest level of armaments will be an essential element in preventing the recurrence of conflict, the Parties within thirty (30) days after this Annex enters into force shall commence negotiations under the auspices of the OSCE to reach early agreement on levels of armaments consistent with this goal. Within thirty (30) days after this Annex enters into force, the Parties shall also commence negotiations on an agreement establishing voluntary limits on military manpower.

According to U.S. Ambassador James Pardew, who helped negotiate the Dayton Agreement and who is the State Department’s Special Representative for Military Stabilization in Yugoslavia, the parties to the accord understood that a private company would train and equip the Federation Armed Forces (FAF), comprised of Croats and Bosnians, in order to achieve the balance of power envisioned in the agreement. Three U.S. companies, SAIC, BDM International, and MPRI, bid for the right to train and supply the FAF, a process monitored by the State Department but controlled by the Federation and its unofficial supporters in the United States. In May 1996, the Federation awarded MPRI a one-year, renewable contract to restructure and develop a modern armed force for the newly integrated forces and to introduce U.S. weapons and train the soldiers in their use. Though MPRI denies that it supplies weapons, many observers speculate that MPRI and Cypress International clandestinely work in tandem. Such an association could give MPRI much more leverage in the international security arena.

The Bosnians had to accomplish two goals before the United States would permit any arms shipments or allow MPRI to commence its training: passage of a Federation Defense Law and compliance with the Dayton Accords’ provisions on the removal of all foreign forces. Compliance with the Dayton Accords forced the Sarajevo government to expel all foreign mercenaries, trainers, or volunteers—meaning the Iranian, Afghanistani, Pakistani, and Turkish mujahidin—and to cut military and intelligence ties to Iran.
Concerned with the influence of the Balkans, the United States used MPRI as a political and military tool in promoting its clear interest in Bosnia.

Though the United States sought to affect the balance of military power and help the Federation forces integrate, it also wanted to retain its perceived neutrality as part of the NATO force in Bosnia in order to act as peace broker in the conflict. The U.S. government was cognizant of concerns that European troops on the ground would be jeopardized by the partisan role played by a U.S. company training Federation forces with Western weaponry. Some Pentagon officials voiced concerns over endangering the U.S. troops still engaged in peacekeeping on the ground in Bosnia. With these concerns in mind, the United States recommended the use of a private company for a role usually reserved for the U.S. military: restructuring the Federation forces into a modern, professional fighting force. Without the badge of state action, the U.S. government hoped to mitigate many of the negative effects of the perception of U.S. favoritism toward the FAF. Consequently, MPRI is currently training and advising the Bosnian Federation.

Some observers claim that MPRI has entered Angola in the wake of EO’s departure, due to U.S. pressure on the Angolan government, and has replaced EO as the leading SC in Angola. However, MPRI denies doing any work in Angola.

In the case of Sri Lanka, MPRI engaged in negotiations with the Sri Lankan government to train a Sri Lankan elite commando unit, to professionalize the military’s command structure, and to assist in NCO training within the Sri Lankan Army (SLA). Since 1983 the internationally recognized government, dominated by the Sinhalese Buddhist majority, has battled the northern rebellion of the Tamil Hindu/Catholic minority, led by the powerful and well-supported Liberation Tigers of Tamil Eelam (the “Tamil Tigers” or LTTE). In 1996, the Tamil Tigers launched devastating offenses against the SLA, prompting the Sri Lankan government to search for additional military expertise, especially in training its navy’s special forces commandos. MPRI claims that the Sri Lankan government withdrew from negotiations with the company for such training and command restructuring.

A State Department source claims that the United States originally approved the license application for the Sri Lankan project because it comported with existing U.S. policy concerning neutrality in the ongoing civil war, i.e., that the United States would not provide the government with “lethal training.” It seems that the U.S. government withdrew its support from the project—thereby extinguishing the negotiations—only when its approval of the MPRI contract had been perceived as a change in U.S. policy toward the civil war and had attracted a great deal of media attention just prior to the November 1996 U.S. presidential elections. It also appeared that the SLA and the LTTE would construe the contract as clear evidence of U.S. support for the regime.

In general, U.S. security companies must obtain official U.S. government approval through the licensing procedure in order to work for foreign governments. Regardless of the amount of direct control or contact by the U.S. State Department or the Pentagon, U.S. security companies act as agents of U.S. foreign policy under the guise of private enterprise in addition to working for their employing countries.
C. Other Companies Worldwide

There are several other SCs which have been operating in or which are being formed in countries that have a recognized level of military expertise and an oversupply of trained military professionals willing to advise, train, and fight for contracting governments or companies. In most cases, the exporting countries sanction the activities of their SCs either tacitly, through clandestine promotion, or officially, through licensing procedures.

In Great Britain, several SCs offer the services of former British special services agents and Gurkha soldiers. Defense Systems Ltd. (DSL), founded in 1982 by former Special Air Service (SAS) commander Alastair Morrison, offers a range of services through the promotion of the British Foreign Office. DSL offers consultation to embassy security guards around the world, guards oil and mining installations in Angola, and trains counter-terrorism units in Indonesia and the Philippines and special forces teams in Jordan, Uganda, and Mozambique. DSL expanded its operations in the 1980s by acquiring other ex-SAS security companies such as Intersec and Falconstar.

Another British company, Gurkha Security Guards, offers the services of the legendary Gurkha fighters used by the British and Indian armed forces. Faced with rebellion and its own military ineptitude, the Sierra Leone government hired Gurkha Security Guards to train its army and to fight on its behalf. The Gurkhas suffered severe losses and were ineffectual against the RUF rebels. The British company was later replaced by the South African company EO.

The Israeli government licenses Israeli SCs to work for legitimate foreign governments and uses the services of such companies as a bargaining chip in negotiations with these governments. For example, after having reestablished diplomatic ties with the government of the Congo in 1991, Israel signed a military agreement in December 1993. As part of that agreement, Israel approved a $50 million contract signed by the Israeli SC Levdan to train Congo President Pascal Lissouba’s palace guard and to supply basic military equipment. Levdan is an Israeli SC and arms firm comprised of retired Israeli soldiers who have served in Lebanon and the Occupied Territories. Opposition members of the Knesset criticized the proposed contract, claiming that Israel should not export military aid to Africa through a private company. After debates in the Knesset, the Israeli government approved Levdan’s contract. The use of retired Israeli soldiers who are still members of the Israeli reserve forces implicates official Israeli participation in any contract signed by Levdan or any other SC employing Israeli reservists.

Worldwide, other companies provide a multiplicity of services. For instance, COFRAS, a French SC, provides demining services and is currently contracted in Cambodia. IDAS, a Dutch Antilles company, and the German firm International Business Company offer troop training and weaponry.

In the months before the fall of President Mobutu Sese Seko’s regime, the Israeli government, along with the Chinese government, sent advisers to Zaire to train thirteen commando brigades for Mobutu’s tottering regime. The Israelis sent official military staff to complete the project as opposed to licensing Levdan or another SC although the
government’s use of an SC would seem ideal in such a controversial conflict so as to avoid official involvement. Like other SCs that have avoided involvement in Zaire’s escalating civil war, Levdan probably did not want to aid a regime which appeared to lack legitimacy and was fighting a losing war. It is also possible that the cash-strapped Mobutu regime could not afford an SC contract and instead was relying on political favors from allies like Israel to obtain needed military expertise.

D. Conclusion

There are many more SCs throughout the world, generated in the militarily advanced countries and employed by regimes in need of military expertise. Unlike rogue mercenaries, SCs provide military training and services in a quasi-official capacity, although their home states are likely to disavow their “private” activities.

The modern notion of the mercenary is that of “an international blight in their perpetration of acts of violence which ruin human lives, create material losses, hamper economic activity and extend terrorist attacks that have touched off or aggravated conflicts, with often catastrophic results for those affected.” SCs, on the contrary, work with recognized states to professionalize militaries and restore a semblance of public order. Since they are registered companies, SCs are constrained not only by their employing states but by the legal obligations and reporting requirements of their home states, as well as by the desire to remain in the good graces of their home states. Market forces, in combination with the extensive media attention SCs receive, require that they maintain a professional reputation respectful of human rights and of their limited mandates. Furthermore, SCs are reluctant to enter arenas in which their governmental employers’ legitimacy is unclear and where there is a likelihood of many casualties. In short, to the extent that SCs actually engage in offensive maneuvers, they do not represent the same threats that mercenaries do. Where SCs have entered conflicts, the ill-disciplined and vengeful national factions fighting for control of resources have been greater threats to the safety and economic well-being of the population.

SCs fill a void by providing military aid in an age when governments do not have the resources or political will to enter internal skirmishes or civil wars on behalf of recognized regimes. At the same time, poor countries are susceptible to low-level guerrilla wars waged by bandits or disaffected members of society because of the countries’ lack of military preparedness and public order. These regimes cannot create well-trained standing militaries or modernize antiquated forces because of high costs and the potential for military coups by existing military institutions.

In this geopolitical context, SCs supply military professionalism, modernization, and expertise to countries wracked by violence and plagued by social disorder. Given the international opprobrium facing foreigners who contribute directly to internal civil wars or unrest in other countries, SCs legitimize their activities and ensure the survival of their personnel by limiting their involvement to military training. In addition to geopolitical constraints, SCs are controlled by market forces which make it too costly in pecuniary terms to enter conflicts in which they might appear to destabilize governments or
aggravate military conflicts. In general, the emergence of SCs represents a new form of state agency which will persist as long as governments need military aid.

IV. THE INTERNATIONAL LEGAL NORMS REGARDING MERCENARIES

Since the 1960s the international community has tried to deal directly with the perceived scourge of mercenarism in the Third World, especially in post-colonial Africa. Africans struggling to gain their independence and establish stable nation-states felt terrorized by the white mercenaries who exacerbated civil conflicts in the nascent states. To Africans, the mercenaries represented racist, exploitative, colonial interests that reveled in the chaos they created. In response to the gross human rights abuses committed by these rogues and the ignoble pecuniary and racist motives which fueled their participation in foreign battles, the international community reacted.

The customary international law prohibiting the recruitment, use, financing, and training of mercenary troops must be examined in light of the historical context of the resolutions and conventions of the Organization of African Unity (OAU) and the United Nations. This Part examines the legality of SCs in light of the developing customary international law against mercenaries. SCs, as presently constituted, do not fall within the definition of mercenaries and their activities are not prohibited by recognized international norms. The prohibitions against mercenaries were not devised to deal with security corporations employed by recognized regimes. These restrictions were also not meant to supersede a sovereign state’s right to employ foreign personnel to restore order or to provide security within their country. The use of SCs by numerous countries, especially by Nigeria, Angola, and other African nations which have led the charge against the use of mercenaries, further demonstrates that SCs are not illegal under international legal norms.

On February 20, 1997, the U.N. Special Rapporteur on the question of the use of mercenaries, Enrique Bernales Ballesteros, issued a report to the U.N. Economic and Social Council that discusses the emergence of SCs and their effect on the international norms regarding mercenaries. The report is important in that it reveals the tensions and problems associated with applying the law of mercenarism to SCs. The report calls for a united stance against the use of mercenaries and expresses concern about the popularity security companies, particularly Executive Outcomes, which has been most successful and is comprised of former “members of racist and extreme right-wing paramilitary organizations.” The report describes the emergence of SCs in the following manner:

The establishment of companies to sell countries military advisory and training services and security services in return for money and mining and energy concessions, in particular, may involve the recruitment of mercenaries not only for military advisory and training tasks in the countries which conclude contracts with them, but also for assistance to the conventional forces of order and public security in combating armed opposition movements and carrying out tasks which should be performed by the police. Where such direct participation does exist, these companies come to take control of the country’s security and have considerable influence over production and economic, financial and commercial activities. Companies of this kind which market security internationally may
acquire a significant, if not hegemonic, presence in the economic life of the country in which they operate.

As seen in the excerpt above, the report is concerned not necessarily with the type and quality of services provided by SCs (as opposed to those provided by individual mercenaries), but by the amount of influence gained by the SCs over contracting states. Throughout the analysis, the report states that mining and other concessions give SCs hegemonic powers. It fails, however, to explain if and how such hegemonic powers have functioned and how, once an SC contract expires, such influence differs from the power of other large mining and energy interests.

The report is unable to reconcile the putative ban on the use of mercenaries with the right of a state to make internal security decisions and pay for any services in whatever form it deems appropriate. As a means of resolving this tension, the report argues that internal security is a role reserved for the state only and should not be delegated to a foreign private company. The report is concerned with SCs’ lack of accountability, and as a result the report claims that the “responsibility for a country’s internal order and security are peremptory obligations which a State fulfils through its police and armed forces.”

The report bemoans and acknowledges that attitudes toward mercenaries are changing, and it accepts the fact that the existing international law regarding their use is incomplete and uncertain. It confronts the dilemma of the existence of SCs, which appear to be organized bands of mercenaries acting on behalf of capitalist interests, and the international community’s apparent acceptance of them. The report continues by assessing the “new concept” regarding the employment of SCs by legitimate governments:

According to this new concept it would appear that any State is at liberty to buy security services on the international market from organizations composed of persons of various nationalities, united by their function and their ability to control, punish and impose the order desired by the Government which hires them, regardless of the cost of lives, in exchange for the delivery of a portion of its natural resources. Naturally, if this hypothesis is confirmed, mercenarism would no longer be considered as necessarily illicit, illegitimate or illegal; however, concepts such as that of State sovereignty and the obligations of States to respect and guarantee the enjoyment of human rights would be tremendously relativized.

What the report fails to recognize is that SCs are distinct from the post-colonial form of “wild geese” mercenaries, and the laws developed to deal with the disruptive force of individual, rogue mercenaries are not applicable to legally recognized SCs. The report asks rhetorically:

Can it be that the mercenaries’ behaviour is changing so profoundly that they now constitute the rank and file of the personnel recruited by a private company in order to contract with African Governments to provide internal security services, safeguard public order, or even to put an end to internal armed conflicts? Not only
is the behavior of SCs and their employees qualitatively different from that of individual mercenaries, but the laws developed to hinder the use of individual mercenaries for disruptive purposes do not apply to SCs.

Though the international community has based the debate over the definition of “mercenary” on the issue of motivation and the vile effects that base motivations create, the crucial factors distinguishing acceptable and unacceptable types of military services offered by foreigners are transparency and state accountability. The primacy of these factors in determining the legality of SCs and of certain types of mercenaries is reflected in several rhetorical questions presented in the report:

Who will be responsible for any repressive excesses that the security companies may commit against the civilian population, especially where representatives of the political opposition are concerned? Who will take responsibility for any violations of international humanitarian law and of human rights that they may commit? . . . What will be the human rights consequences of entrusting internal order and control over the exercise of civil rights in a country to an international private security firm? Is the international community willing to accept and concur with the idea that the recruitment of mercenaries is illegal only in a few very limited cases? When, and in what circumstances, should the recruitment of mercenaries be considered legal?

The purpose of this section is to help reconcile the apparent tension, engendered by the emergence of SCs, between the ban on the use of mercenaries and the rights of states to provide for their internal security. This Part will show that a total ban on the use of mercenaries does not exist in international customary law and that SCs fall outside those prohibitions that do exist. Therefore, in answer to the final question presented above, there are cases in which the use of mercenaries by recognized regimes or movements of national liberation is legal, and the international community is growing more willing to recognize this distinction.

A. Private International Security Companies as Mercenaries?

One of the major obstacles to restricting the recruitment and use of mercenaries has been defining the “mercenary” and his acts accurately. States have differing interests in defining the term narrowly or liberally. Especially for those countries most involved in the mercenary market, finding an acceptable compromise is essential to the effectiveness of any regulation in the international system, where the subjects of the law often determine the development of the law through their compliance. There are three essential characteristics which distinguish a mercenary from other combatants: (1) the mercenary is not a citizen or resident of the state in which he is fighting; (2) the mercenary is not integrated (for the long term) into the national force; and (3) the mercenary does not have government support for his actions.

One could define a mercenary liberally as a foreigner serving in an organized armed force not his own; however, this definition may be too broad for the accepted practices of the international community. First, the British, French, Indians, the Vatican, and various
countries in the Middle East would object to such a definition since their armed forces include foreign regulars. Second, the definition would include officers and contractors seconded to foreign armed forces, as occurs between Great Britain and certain Gulf states. Third, it would include international volunteers who engage in foreign wars for ideological or political reasons, like the Lincoln Brigade in the Spanish Civil War.

Currently, the mercenary is defined primarily by his motivation: money. Focusing solely on a pecuniary motivation also creates an unworkable standard upon which to build an international legal norm because of the near impossibility of determining the motivations of combatants. The Diplock Report, commissioned in 1976 to study the issue of mercenaries, recognized this problem and concluded that mercenaries “can only be defined by reference to what they do, and not by reference to why they do it.” On one hand, this interpretation captures the heart of the international concern over mercenaries—the potential negative effects of mercenary actions. These include human rights abuses, chaos, aggravation and prolongation of a conflict, and suppression of popular will. On the other hand, defining mercenaries by the effects of their actions is an ex post determination that produces an inconsistent and vague standard.

More importantly, this definition ignores state accountability, the pivotal element that makes a combatant legal or not. The tie between a state and an individual demarcates the dividing line between the criminal acts of a mercenary and the legal acts of a state player (legal acts that can be criticized under different standards). The international community’s fear of mercenaries lies in that they are wholly independent from any constraints built into the nation-state system. The element of accountability is the tacit standard that underlies the international antipathy for mercenary activity and truly determines mercenary status.

Nevertheless, the international community has formally adopted motivation as the determining factor that characterizes a mercenary. In the Additional Protocols to the Geneva Convention of 12 August 1949 (Protocol I, Art. 49), a mercenary is defined as any person who:

(a) is specifically recruited locally or abroad in order to fight in an armed conflict;

(b) does, in fact, take a direct part in the hostilities;

(c) is motivated to take part in the hostilities essentially by the desire for private gain and, in fact, is promised, by or on behalf of a Party to the conflict, material compensation substantially in excess of that promised or paid to combatants of similar ranks and functions in the armed forces of that Party;

(d) is neither a national of a Party to the conflict nor a resident of territory controlled by a Party to the conflict;

(e) is not a member of the armed forces of a Party to the conflict; and

(f) has not been sent by a State which is not a Party to the conflict on official duty as a member of its armed forces.
This definition is tightly constructed so as to exclude foreign nationals in other country’s armed forces that the international community is willing to tolerate. Additionally, Article 47 of Protocol I ignores those foreigners integrated into the armed forces of another state, those who are motivated by ideology or politics, and those who may not actually fight in the hostilities. Trainers and advisors are also excluded, although they can affect the military situation just as much as, if not more than, actual combatants. In the case of SCs, this is a crucial omission since most of the work SCs do involves training. The comments to Article 47 of Protocol I further note that foreign advisors and military technicians are prevalent throughout the world and are necessary given the technical character of modern weaponry. The International Convention against the Recruitment, Use, Financing, and Training of Mercenaries (“International Convention”), which is not in force, adds to the Protocol I definition by including those persons who recruit, use, finance, or train mercenaries as offenders. Though both definitions contain the element of motivation, they are primarily concerned with private acts of violence uncontrolled by any state.

Those who object to companies like EO claim that SCs fall under the Article 47 definition of mercenaries: The employees of SCs are foreign military soldiers specially recruited abroad and paid in excess of what the military personnel of the contracting state are paid; they engage in fighting; and they are not sent on official duty of the armed forces of another state.

SCs in general, however, fall outside the conjunctive definition of Article 47 since they tend to restrict their activities to training government troops only. Assuming, arguendo, that all SCs are like EO, which has in the past engaged in offensive maneuvers, the definition includes additional distinctions. First, section 1(a) requires that the combatant has to be recruited to “fight” in an “armed conflict.” If SCs are not contracted to “fight,” and do so only in defense, then SCs would fall outside the purview of the definition. Furthermore, the definition of a situation of “armed conflict” is a subjective determination, and SCs’ contracts could specify that the agreement involves the restoration of security as opposed to participation in an armed conflict. This could also support the argument that SCs fall outside the Article 47 definition. Second, most SCs integrate into the armed forces for which they work, as EO did in Sierra Leone, thereby fulfilling the requirement in 1(e). According to the comments to Article 47, “it is clear that enlistment in itself is sufficient to prevent the definition being met.” Third, it could be argued that an SC, which receives authorization from its home state to operate abroad (through a licensing process or more informally), is “sent by a State which is not a party to the conflict on official duty” and that the SC represents a member of that state’s armed forces. Under an extreme interpretation of section 1(f), SCs could be characterized as civilian contractors, although regarded as a member of the military force in the field. This implies that SCs can be regarded as a contractor of their home states or of their employing states. In either case, SCs would be tied to state actors and would fall outside the definition of a mercenary in this aspect as well.

Even beyond the provisions of Article 47, it is clear that SCs are not mercenaries, particularly because state accountability is the key to distinguishing mercenaries from other combatants. SCs are tied to states in various unofficial ways. First, SCs are legal entities, bound to employing states by recognized contracts and to home states by laws
requiring registration, periodic reporting, and licensing of foreign contracts in most cases. Second, most SCs are willing to share information with their home governments because the home governments tend to be large, repeat customers that appreciate voluntary cooperation. Third, because of the importance of the home state, SCs try to act in their home states’ interest abroad, thereby making the SCs quasi-state agents. Fourth, since SCs are long-term market players (unlike most individual mercenaries), they are unlikely to engage in wanton violence that potentially could spark opposition in employing states or their populations, home states, or future customers.

All of these relationships presuppose that the home state has some control over the existence and profitability of SCs. At present, this control is tenuous in the case of South Africa’s EO, although the ANC is proposing new legislation to restrain EO’s activities. With some home-state control over the profitability of the SC, the SC is linked inextricably with the state.

For these reasons, SCs are not mercenaries. Not only do SCs fall outside the technical definition of a mercenary, but their activities are not prohibited “in light of other provisions relating to mercenaries and the restrictive approach adopted in various United Nations resolutions which link mercenaries with concerted acts of violence aimed at violating the right of peoples to self-determination and undermining the constitutional order of a State or its territorial integrity. . . .” The sine qua non of mercenarism, therefore, is not motivation or nationality, but the purpose for which a mercenary is employed. Often, the legitimacy of the mercenary’s purpose is tied directly to the legitimacy of his employer. However, the mercenary can only exist if he violates the accepted norms of the state-based system. . . .

B. Development of a Norm Against the Use of Mercenaries

The genesis of the international movement to ban the use of mercenaries followed the unsettling effects of mercenary activities in post-colonial Africa. Mercenaries were neocolonial soldier-bandits who helped subvert national movements of liberation against colonial regimes and served as instruments of foreign aggression against nascent regimes. The resolutions and conventions banning the recruitment and use of mercenaries developed in the context of mercenaries who challenged the sovereignty of states and the independence of movements of national liberation.

Before 1945, the international community considered it an act of belligerence for a state to permit the recruitment or enlistment of mercenaries to attack another state. Article 2(4) of the U.N. Charter prohibits member states from utilizing “the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations.” The international community extended this obligation to include the duty of states to prevent the organization of bands or mercenary troops intending to attack the sovereignty or territorial integrity of another state. According to Chapter I of the Convention (V) Respecting the Rights and Duties of Neutral Powers and Persons in Case of War on Land signed in 1907 at the Hague, neutral states during international wars cannot allow “corps
of combatants” or “recruiting agencies opened on the territory of the neutral Power” to aid the belligerents.

The Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States extended this duty in accordance with the Charter of the United Nations adopted in 1970: “Every State has the duty to refrain from organizing or encouraging the organization of irregular forces or armed bands, including mercenaries, for incursion into the territory of another State.” Other General Assembly resolutions reaffirmed this obligation of states not to allow, either proactively or through inaction, organized groups to invade other countries. Even so, states do not have affirmative obligations under traditional notions of international law to stop individuals from exiting their respective countries to join mercenary groups. From these traditional notions of neutrality, the international community has developed norms prohibiting the use of mercenaries.

In the modern international environment, restrictions on the use of mercenaries developed within colonial Africa. During the Katangan secession in the Congo, the Security Council passed several resolutions calling for the exit of all foreign personnel and sent U.N. troops, under the United Nations Operation in the Congo (ONUC), to combat the white mercenaries fighting on behalf of Katanga. The United Nations finally crushed the secession in 1963 and departed in 1964. Following the grave impact of Bob Denard, “Mad” Mike Hoare and other mercenaries in the Congo’s numerous internal conflicts and the appearance of mercenaries in other countries like Nigeria, the OAU addressed this problem and began to codify the prohibition against mercenaries.

On several occasions in the 1960s and 1970s, the OAU passed resolutions condemning the presence of mercenaries in various conflicts throughout the continent. In Addis Ababa in 1971, the Assembly of Heads of State and Government of the OAU declared that mercenaries represented a threat to the “independence, sovereignty, territorial integrity and the harmonious development of Member States of the OAU” and condemned the use of mercenaries to jeopardize the sovereignty of member states. In 1972, the OAU’s Council of Ministers’ Committee of Legal Experts drafted the Convention for the Elimination of Mercenaries in Africa, which the OAU finally signed in Libreville in 1977. This convention became the seminal codification of an international ban on mercenaries. Article 1 of the convention defined the mercenary by referring to the purpose of his employment. The mercenary’s acts were also declared crimes against the peace and security of Africa:

[A] “mercenary” is classified as anyone who, not a national of the state against which his actions are directed, is employed, enrolls or links himself willingly to a person, group or organization whose aim is:

(a) to overthrow by force of arms or by any other means the government of that Member State of the Organization of African Unity;

(b) to undermine the independence, territorial integrity or normal working of the institutions of the said State;

818
Draft — do not distribute!
(c) to block by any means the activities of any liberation movement recognized by the Organization of African Unity.

However, the OAU convention “does not explicitly forbid the employment of mercenaries.” Article I prohibits governments from hiring mercenaries to suppress movements of national liberation, but the African countries carefully constructed the convention to allow legitimate governments to defend themselves from “dissident groups within their own borders” by hiring non-nationals. The major concern for African countries at this point was that mercenaries not be used against OAU-recognized liberation movements.

Although the colonial powers tried to avoid discussion of the issue, the United Nations (goaded by Second and Third World countries) was unable to avoid the topic since mercenaries were engaged in several countries and their acts were notoriously brutal. The United Nations followed the OAU’s lead and passed several resolutions calling for measures to end the recruitment and use of mercenaries in a number of conflicts.

Throughout the late 1960s and 1970s, the Security Council responded to specific crises by calling for the removal of all mercenaries. For example, the Security Council, on several occasions, condemned Portugal for allowing mercenaries to operate from its colonial territory. The General Assembly also passed numerous resolutions condemning the use of mercenaries for the destabilization of a regime or against a movement for national liberation. In 1968, the General Assembly declared:

The practice of using mercenaries against movements for national liberation and independence is punishable as a criminal act and [ ] the mercenaries themselves are outlaws, and [we call] upon the Governments of all countries to enact legislation declaring the recruitment, financing and training of mercenaries in their territory to be a punishable offence and prohibiting their nationals from serving as mercenaries.

The General Assembly followed this proclamation of criminality five years later by condemning the use of mercenaries by “colonial and racist regimes against the national liberation movements” as criminal. This resolution also gave movements of national liberation (groups fighting racist or colonial regimes) international status in terms of the Geneva Conventions of 1949 relating to the rights and duties of combatants. The United Nations, however, soon realized that its sporadic pronouncements condemning the use of mercenaries in particular situations were not having the desired effect because states failed to restrain their citizens from enlisting in mercenary groups.

As a result, the Diplomatic Conference included the problem of mercenarism in its proposed amendments to the Geneva Conventions of August 12, 1949. On June 8, 1977, the Diplomatic Conference adopted the Additional Protocols I and II to the Geneva Conventions by consensus. Article 47 of Protocol I defines a mercenary and strips all mercenaries of the right to claim prisoner-of-war status, a right presumed under Article 45(1) of Protocol I. Under this Protocol, states may still offer mercenaries prisoner-of-
war status if they choose, but after 1977, mercenaries have no right to combatant status and can be tried as common criminals in the offended state.

Countries have used U.N. and OAU resolutions to prosecute mercenaries. The most visible trial occurred in Angola where thirteen white combatants were tried for the crime of mercenarism. In 1976, the Angolan government (the victorious MPLA) charged the mercenaries based on OAU, U.N., and other international resolutions although Angola at the time did not have a law prohibiting mercenarism. The Angolan courts sentenced nine of the mercenaries to prison terms and condemned the other four to death. The Angolan trial led to the drafting of the Luanda Convention on the Prevention and Suppression of Mercenaries by invited observers of the trial. Among other stipulations, the Luanda Convention stripped mercenaries of combatant status and formed the basis for parallel provisions in Additional Protocol I. These trials indicate that the international prohibitions reflected in the resolution have been accepted as universally applicable and binding on all member states.

The continued use of mercenaries, despite the United Nations’ constant protests, forced the international community to recognize the need for a multilateral convention. In 1980, the General Assembly resolved to include the drafting of an “International Convention against the Recruitment, Use, Financing and Training of Mercenaries” in the provisional agenda for the thirty-fifth session. During the session, the General Assembly formed a committee to draft the international convention. The committee spent several years assessing respective state proposals and developing the treaty. On December 4, 1989, the General Assembly presented the International Convention for signature and ratification.

The International Convention establishes a more expansive definition of mercenary yet still relies on motivation to distinguish mercenaries from other types of combatants. The recruitment, use, financing, and training of mercenaries are all offenses under the convention, and states are required to prevent these offenses. Although it appears to crystallize the customary international law regarding mercenaries, three main criticisms of the International Convention exist. First, states are given jurisdiction for the crime of mercenarism only when offenses are committed in their territories or by one of their nationals. Second, the International Convention fails to address the treatment of states in conflict by not allowing aggrieved states to proceed against offending states. Third, there is no monitoring mechanism, leaving implementation in the hands of the states.

In addition to the International Convention, the United Nations has continued to issue resolutions since the 1980s. These resolutions reflect the limited nature of the ban on the use of mercenaries and the traditional concerns the international community regarding individual mercenaries. These resolutions deal with the following uses of mercenaries: (1) to destabilize neighboring states, as in the case of South Africa in the 1980s; (2) to spearhead coups in small states like the Seychelles; (3) to prevent the granting of independence to movements of national liberation; and (4) to violate human rights. In essence, the United Nations recognized in these resolutions that the activities of mercenaries are contrary to the fundamental principles of international law, such as “non-interference in the internal affairs of states” and “territorial integrity and independence.” Marie-France Major describes this type of mercenarism as an “internationally wrongful
act.” The historical development of these norms supports the notion that only internationally wrongful acts are proscribed with respect to mercenary activity. The international community has only witnessed mercenaries in their destructive, rogue form and consequently has developed norms to deal with those particular types of activities.

Some scholars argue that the Security Council’s repeated appeals for governments to restrict the supply and demand for mercenaries is evidence that a “state has an obligation [that goes beyond the traditional constraints of international law] to control the recruitment of its nationals in situations where a threat to peace and security exists.” However, the argument overlooks the fact that the Security Council resolutions addressed particular conflicts where the Security Council judged that mercenaries were aggravating conflicts and presenting threats to international peace and security. Much of the discussion also assumes that the General Assembly resolutions, which are broader in scope, reflect established customary international norms. Yet under the U.N. Charter, the General Assembly has no authority to enact, alter, or terminate rules of international law. Therefore, General Assembly resolutions do not necessarily constitute international law. Instead, resolutions from the General Assembly or regional organizations like the OAU may only represent the crystallization of customary international law or evidence of state practice and opinio juris.

Furthermore, the International Convention has not been ratified by enough states for it to be in force. Given the tenuous nature of the definitive convention, it is difficult to argue that these norms extend to SCs. Of the sixteen signatories, three—Angola, the Congo, and Nigeria—have hired or dealt directly with SCs, while another, the Ukraine, is a major supplier of pilots to EO. Zaire, another signatory, has hired individual mercenaries to fight the powerful belligerents led by Laurent-Desire Kabila. It is difficult, therefore, to claim that a standard exists beyond that inherent in the peremptory norms of international law.

C. State Practice Regarding Mercenaries

The status of customary international law banning the use of mercenaries is unclear because of the lack of consistent state practice and evidence of opinio juris. There appears to be an international antipathy, evidenced in numerous U.N. resolutions and reports, toward the use of mercenaries to destabilize legitimate governments or to suppress recognized movements of national liberation. However, the International Convention still has not entered into force, and “the laws of most countries do not make mercenary activities a criminal offence.” Many countries have regulations that restrict the recruitment, training, and use of mercenaries within their national territories, but most countries do not prohibit their citizens from traveling abroad and enlisting as combatants. These domestic legal positions reflect individual countries’ concerns that private citizens will implicate their home states by violating the sovereignty of other states. But most states have neither comprehensive legislation prohibiting mercenary activity by their citizens nor vigorous enforcement of the laws that do exist.

As international concern about the implications of private military force has grown, states have adopted neutrality laws that restrict private acts of violence. By the mid-nineteenth
century, one-third of all states had enacted restrictions on foreign military service. Current laws in most countries still reflect the principal objective of preventing state responsibility for the actions of private citizens. As a result, there are powerful proscriptions against mercenary activity which could jeopardize the neutrality or interests of home states, and less concentration on the absolute abolition of mercenary activity.

1. The United States

The United States has passed several legislative acts restricting the potential recruitment and enlistment of its citizens as mercenaries for foreign agents: the Neutrality Act of 1794; the Foreign Relations Act; the Immigration and Nationality Act; and the Foreign Agents Registration Act. The U.S. government first addressed the problem of private citizens launching hostile military expeditions in the Neutrality Act of 1794 (“1794 Act”). The 1794 Act was passed during George Washington’s administration to prevent private citizens’ engagement in the Anglo-French War and the possibility of U.S. involvement. U.S. laws have undergone several transformations in response to various international conflicts in which U.S. citizens were involved, such as the Cuban rebellion against Spain and the Spanish Civil War.

The Foreign Relations Act regulates those activities of private citizens which may jeopardize U.S. foreign policy interests. Section 959 prohibits any person in the United States from enlisting, recruiting, or leaving U.S. territory in order to serve any foreign entity or agent in a military capacity. The key element that triggers the Foreign Relations Act is that the proscribed activity be conducted within U.S. territory. With respect to the crime of leaving U.S. territory with the intent of serving in a military capacity, the Supreme Court ruled that the government has no power to prevent its citizens from joining foreign armed forces when such acts are executed outside U.S. territory.

The Immigration and Nationality Act contains a provision which revokes the U.S. citizenship of any individual who undertakes certain activities “with the intention of relinquishing United States nationality.” Under section 1481(a)(3), a U.S. citizen may be stripped of his citizenship for serving in a foreign armed force:

A person who is a national of the United States whether by birth or naturalization, shall lose his nationality by voluntarily . . . entering, or serving in, the armed forces of a foreign state if (A) such armed forces are engaged in hostilities against the United States, or (B) such persons serve as a commissioned or noncommissioned officer.

This language reflects a concern on the part of the U.S. government to ensure that its international relations are not complicated by the independent acts of its citizens. Yet, the constitutionality of this provision has been called into question by the Supreme Court’s ruling in Afroyim v. Rusk. In Afroyim, the Court held that under the Due Process Clause of the Fourteenth Amendment the government cannot divest a citizen of his citizenship for simply voting in a foreign election. The Court’s general proposition that a U.S. citizen has “a constitutional right to remain a citizen in a free country unless he voluntarily relinquishes that citizenship,” arguably could be extended to invalidate this subsection of
the Immigration and Nationality Act. To date, no American citizen has been divested of his citizenship under this subsection.

Finally, the Foreign Agents Registration Act requires any agent of a foreign state to register with the federal government, including agents recruiting military personnel. Failure to register properly creates criminal liability; hence this law allows the government to control the recruitment of U.S. citizens or residents by foreign nations.

Allaoua Layeb argues that U.S. legislation is ineffective because it was designed to restrict U.S. citizen enlistment in foreign armed forces. The laws are not geared to restrict the use of modern mercenaries by interest groups or individuals fomenting conflict in other states. In general, “the laws are not aimed at preventing an individual from leaving the country with the intent of enlisting in a foreign army.” Instead, “[t]he real purpose of these neutrality laws, is to prohibit the commission of unauthorized acts of war by private United States citizens because these unauthorized acts may lead to acts of reprisals against the whole nation, as well as to civil disunion.”

Furthermore, “[t]he United States Government has long pursued a policy of non-enforcement with regard to its neutrality laws,” as evidenced by the independent ventures of U.S. citizens in both world wars, the Spanish Civil War, the Cuban crisis, and the Angolan conflict. After the Angolan trial of thirteen mercenaries, which resulted in the execution of one American citizen and four British subjects, the House of Representatives Foreign Relations Committee held hearings to discuss U.S. legislation and the recruitment of U.S. citizens to fight abroad. The Assistant Secretary of State for African Affairs denied U.S. government involvement in the recruitment of mercenaries to Angola and also denied that mercenarism (at the time) was a crime in international law. The Angolan experience did little to invigorate U.S. enforcement of its neutrality laws against mercenaries. The U.S. government has utilized its laws only on a few occasions which involved radical filibuster campaigns. In May 1981, ten white supremacists were stopped from invading Dominica in the “Bayou of Pigs” event. The U.S. also arrested fourteen mercenaries, led by former U.S. Customs Service agent Tommy Lynn Denley, who planned to overthrow the regime of Lieutenant Colonel Desi Bouterse in Suriname.

In short, U.S. laws are incomplete and ineffectively administered. This shows a lack of political will on the part of the United States to condemn all mercenary activity given that the development of a total ban on mercenary activity would put U.S. citizens in danger of prosecution abroad. It is no surprise, therefore, that the U.S. government has not prosecuted U.S. security companies under these laws and does not consider SCs to be mercenary organizations. Other countries have followed a similar pattern.

2. The United Kingdom

As in the United States, the Angolan mercenary trials, in which ten British subjects were tried and three were executed, forced the British government to evaluate its antimercenary laws. In the Diplock Report, the Committee of Privy Counsellors appointed to inquire into the recruitment of mercenaries discussed the faults with its Foreign Enlistment Act of 1870 (“1870 Act”) and the limitations of international laws
restricting the use of mercenaries. The 1870 Act prohibits British subjects from recruiting for or enlisting in the armed forces of a “foreign state.” It also creates a crime for leaving British territory by ship with the intent of accepting an illegal commission. There are exceptions to these prohibitions, such as government licenses, and anachronistic loopholes, such as the absence in the statute of “airplanes” as a means of transport or attack. In general, a British citizen may be prosecuted only for acts committed within British jurisdiction.

The Diplock Report concluded that the international definition of “mercenary” based on the motivation of the combatant was unworkable. The Report concluded that “[t]o serve as a mercenary is not an offence under international law.” As a result, the Privy Counsellors resolved that preventing British citizens from working abroad as mercenaries was an unjustified infringement on individuals’ personal freedom. The report advocated the repeal of the unworkable provisions of the 1870 Act, and advised that any new legislation permit enlistment and service as mercenaries abroad. It recommended a legislative system in which the government would have the power to publish a list of those armed forces in which British citizens could not enlist. This legal regime would allow the government to determine what effect particular mercenary enlistments would have on Great Britain’s international relations.

Since the issuance of the Diplock Report, the government has made no legal or practical changes. In the late 1980s, British mercenaries were employed by drug cartels, but the limitations of British law hindered the prosecution of these individuals. Great Britain remains the center of mercenary recruitment in the world. EO and Sandline International along with a host of British SCs are registered in Great Britain. British SCs operate freely under the licensing provisions of the British government. . . .

4. Other Countries

Other countries have passed minimalist legislation and have failed to implement the legislation that does exist. France, for instance, outlaws the recruitment of mercenaries on behalf of a foreign power under Article 85 of its Penal Code, but it rarely enforces this policy as evidenced in recent forays by French mercenaries in Zaire. It also allows SCs like COFRAS to operate relatively freely. The Ukraine and the Slovak Republic outlaw their citizens from serving as mercenaries without the approval of the appropriate government agencies. In short, many countries outlaw the recruitment of their nationals to serve abroad without their governments’ consent—negating the idea that states accept an absolute prohibition against the use of mercenaries.

5. State Practice Toward SCs

State practice regarding mercenaries demonstrates that there is a general acceptance of the phenomenon of mercenarism while many states outlaw the recruitment or use of mercenaries within their own territories. Even regional organizations like the OAU have contributed to this development. The OAU Convention, the most aggressive codification of the criminality of mercenarism, envisioned the legitimate use of foreign military personnel to support regimes. Its own policy to support incumbent regimes facing
internal challenges highlights the principle that the legitimate state is entitled to the use of foreign military expertise. The new norm developing with respect to SCs is that legally registered companies that provide security services for struggling yet legitimate regimes are not mercenaries.

For instance, the government of Angola finds no inconsistency between its previous prosecution of mercenaries and its employment of SCs. It does not consider EO to be a mercenary organization, and it lauds EO for its devoted and professional service. . . . Furthermore, other countries like Nigeria and Ghana have worked with EO to achieve mutual security goals. Humanitarian organizations have also utilized SCs’ services, finding them to be effective, objective, and trustworthy. Even the OAU, the bastion of antimercenarism and the veritable font of antimercenary legislation, has leveled only muted criticism against EO and has even considered employing the firm for continent-wide peacekeeping.

Perhaps the persistent use of mercenaries and lax enforcement of existent prohibitions on their use reflect the utility of mercenaries in certain contexts: “Whatever the laws, whatever the international conventions, whatever the deterrents, mercenary activities will continue as long as governments or counter-governments see a use for mercenaries and as long as would-be mercenaries are themselves ready to be used.” SCs seem to be an even more acceptable form of private military force that can be contained by market forces and state regulation. For this reason, state practice demonstrates that states consider SCs to be legitimate organizations performing useful services in the international arena. Even if it is argued that state practice historically reflects opinio juris for banning the use of mercenaries, the treatment of SCs as legal entities is evidence of a different attitude toward these organizations and a shift away from an absolute abolition of the use of independent foreign military personnel.

D. The Inherent Tension Between State Sovereignty and an Absolute Ban on the Use of Mercenaries

The development of the customary international law banning mercenaries cannot override the peremptory norms of international law. Two such norms clash directly with a ban on the use of mercenaries which prohibits recognized, legitimate states from employing private military expertise. First, Article 51 of the Charter reserves to the state the right to self-defense: “Nothing in this present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations.” Second, Article 2(7) restricts the authority of the United Nations to meddle in a state’s internal affairs: “Nothing contained in the present Charter shall authorize the United Nations to intervene in matters which are essentially within the domestic jurisdiction of any state.”

Article 3 of Protocol II expresses this limitation on the degree to which customary laws developed in treaties and resolutions may supersede peremptory norms regarding the rights of the state:
Nothing in this Protocol shall be invoked for the purpose of affecting the sovereignty of a State or the responsibility of the government, by all legitimate means, to maintain or re-establish law and order in the State or to defend the national unity and territorial integrity of the State.

In addition, the International Convention is necessarily restricted to the “principles of international law such as those of sovereign equality, political independence, territorial integrity of States and self-determination of people.” The operative term in Article 3’s statement of nonintervention is “legitimate means.” The question must, therefore, be raised whether the use of mercenaries or SCs in particular situations on behalf of a legitimate regime facing internal unrest is a legitimate means of restoring internal order.

The right to self-defense in a classical international conflict entitles the aggrieved state to defend itself in any way possible until the Security Council has taken “measures necessary to maintain international peace and security.” In the case of a state that does not have recourse to a prepared standing army, it may be necessary to enlist the aid of allies or private military professionals. Usually the latter option is not used either because allies are willing to help or because there is no private military expertise available to repulse the invaders.

The right may also arise in an internal conflict in which the armed rebellion or insurgency is supported and aided by a foreign power. In the case of Sierra Leone, the RUF had the military, political, and strategic help of both Libya and Liberia. In such a scenario, Sierra Leone would have been justified in hiring mercenaries or an SC to help it repel the attacks and restore internal order. (Of course, the employing state may not abdicate its responsibility to supervise the activities of the mercenaries or SC it hires.) Tanzania declared the right to request assistance: “[A]ny African state has the right to ask for assistance, either military or economic, from the country of its choice... [and it is] surely justified in seeking assistance to overcome a temporary crisis, and the donor country should not be accused of neo-colonialism for responding.” Certainly that aid can come in whatever form as long as the aid does not violate other established norms, like human rights and sovereignty principles.

The traditional maxim in international law is that whatever action the law does not expressly require or prohibit is left to the states. Even so, the increase in the number of small-scale wars and their spill-over effects, seen most graphically in the Great Lakes region of Africa, have forced the international community to view internal conflicts as potential international threats to peace and security. As a result, it is anachronistic and overly simplistic to state that regimes may do whatever they please within their borders. Regimes that are militarily challenged internally are constrained when their opponents reach the recognizable status of belligerents.

The OAU, however, has a standard policy to support its member governments in the case of civil wars, as opposed to remaining neutral. If a resistance movement has not attained belligerent status or is not a movement of national liberation, then it is arguable that the incumbent regime may employ foreigners to help it regain control of the country. The OAU would certainly support the established regime, as long as it were not engaging in
illegal practices in restoring its control. When it hired EO, Angola’s democratically elected MPLA government acted within its right to restore order internally. The OAU offered restrained criticism and later began negotiations with EO to provide security services for its organization.

Yet, SCs could be seen as intervening agents of foreign states who are interfering with the internal political convulsions of other states. Arguably, SCs could then be seen as illegal per se if they interfere with the right of people to self-determination. But if SCs work for democratically elected or U.N.-recognized governments to suppress civil unrest, they cannot be characterized as illegitimate intervenors. In those cases, they are contractors for employing states and the vicarious agents of the popular will.

In general, a total ban on the use of mercenaries is inconsistent with superseding norms of international law. Recognized states may hire mercenaries or SCs to defend themselves from external attacks or to restore order internally. As long as these regimes do not violate other norms in the employment of foreign military expertise, they are justified under international law to avail themselves of the military aid they require.

\[E. \text{Conclusion}\]

It is ironic that states have needed to resort to private sources of military force to restore their own sovereignty. By delegating certain military and security duties to private companies or outsiders, states are abdicating responsibilities that are normally reserved for the state. Yet in so doing, they are injecting objective military strength, which is often what is lacking in chaotic situations. Much like the dukes who employed *condottieri* in Renaissance Italy, states like Sierra Leone are searching for experienced military professionals who will conduct their missions objectively and efficiently and will leave when they are told.

The international law regarding mercenaries does not inhibit the right of states to employ mercenaries or SCs for legitimate ends, like restoring social order or defending against external aggression. Further, SCs do not fall within the international proscriptions banning the use or recruitment of mercenaries. SCs represent quasi-state actors in the international arena, which takes them outside the mercenary concerns of the international community. SCs can pose a danger if they are taken out of the state-controlled system; however, as presently operating, SCs are not mercenaries.

\[V. \text{The Normative Considerations for the Existence of SCs}\]

Although the current customary international law regarding mercenaries does not apply to SCs, the international community needs to determine if SCs should be regulated. . . .

Since the 1960s, independent mercenaries have gained the reputation of being amoral, ruthless thugs who merely aggravate conflicts, threaten struggling regimes, commit human rights atrocities, and serve the interests of colonial or capitalist entities. The international community and some national legislators have responded to this perception. Now, however, SCs have emerged, and countries are trying to understand their character and motives. The [UN] Special Rapporteur[, in his 1997 report on mercenaries,] has
recognized that the “mercenary” market is changing, but he has concluded that the changes have created a greater danger for state sovereignty and international peace and security. As a result, he has called on the international community to ban mercenary activity in all its forms, including SCs. Unfortunately, his conclusions about the new “dangers” are incomplete because he has failed to analyze the nature and dynamics of the international market forces to which the SCs are responding.

. . . In general, he believes that SCs are a “threat to the self-determination of peoples and an obstacle to the enjoyment of human rights by peoples who have to endure their presence,” and a threat to state sovereignty. In order to include SCs as “mercenaries,” the report expands the traditional definition of these “crimes” with the following:

(1) a “threat to self-determination” does not simply signify interference with a movement of national liberation or with a national election, but an intervention of any sort in the internal affairs of a state, regardless of the threat faced by the government or the population;

(2) an “obstacle to the enjoyment of human rights” means any use of armed force, regardless of how it is used or against whom it is directed; and

(3) a “threat to state sovereignty” means the assumption by foreigners of any duties previously reserved to the state, not just an attack on the legitimate government of the state.

A. Negative Consequences

There are several potential negative consequences from the emergence of SCs. To the Special Rapporteur, EO presents the best example of the new threats posed by SCs.

First, EO is not accountable to any state since it has no formal connections with the South African government.

Second, EO is comprised of ex-members of the most notorious and ruthless commando units of the apartheid regime, used in “mercenary” fashion to attack the sovereignty of southern African countries. In the past, these individuals acted without concerns of accountability or human rights.

Third, the company engages in actual combat, raising questions of accountability in the field, undue influence on the internal conflicts of a state, and potential human rights abuses. EO’s involvement in the field also raises the possibility that it will act solely for its own benefit against all actors (including EO’s employers) in the conflict, much as Denard and Schramme led their infamous mercenary revolt in the Congo in 1967.

Fourth, EO and other SCs are closely tied to other companies with diverse economic interests, such as arms dealers, energy and mining companies, and construction firms. As a result, EO and other companies appear to be engaging in African and other conflicts solely as a means of obtaining concessions and related contracts for their corporate brethren. By aiding struggling regimes willing to grant valuable concessions, EO and its
parent SRC could then gain hegemonic power in the countries, leading a vanguard for the “neocolonialism of the twenty-first century.” Not only does this type of corporate association represent an amalgam of concentrated economic and military power reminiscent of the mercantile companies of the eighteenth century, but it creates the opportunity for greater corruption and influence on the internal dynamics of the employing states.

Fifth, SCs could fuse their power with that of arms traffickers, drug dealers, and terrorist groups, thereby creating an unholy alliance of non-state agents with the economic, military, and political power to overwhelm states and the state system in general. They could also assist rogue states unable to receive military aid through the international state system.

Sixth, SCs could switch sides in an ongoing conflict based on the highest bidder.

Seventh, SCs could turn on their home state and represent a challenge to the legitimate use of force in that society. In the least, SCs are competitors for the best military talent in the home state.

Eighth, SCs could aggravate conflicts and give military life to illegitimate regimes that would otherwise have to settle their conflicts peacefully. In this sense, SCs could be agents of the status quo, aiding only those governments with enough money to retain power while suppressing potentially legitimate resistance movements.

Ninth, the supply of SCs gives incumbent regimes an incentive not to prepare their defense forces properly and to cede their internal security duties to private agents.

Tenth, SCs could fight each other or their own countries’ national forces in third-party countries. Even if this scenario does not appear likely, there is the possibility that SCs could drag their home states into conflicts or force them to come to the SCs’ assistance with national military power. Also, former clients could use the skills learned from SCs against SCs’ home states.

Eleventh, SCs may allow countries to supersede public debate about involvement in foreign countries by subcontracting their foreign policy to private companies.

Twelfth, SCs also could be used as covert agents.

Finally, SCs simply legitimate the profession of mercenarism, thereby unleashing the dangerous threat of mercenaries in general.

All of these negative effects represent a potential challenge to the nation-state system and the traditional powers of the state.

B. Market Forces

These problems, however, are rectified to a great extent by state regulation and the market itself. SCs are legal, registered companies often subject to some form of national
regulation. In many countries, the government must license each SC contract, thereby imposing some form of state regulation and accountability for the terms of employment. This reduces the potential for conflicts between the home state and the SC. SCs are profit-making organizations with an interest in appearing legitimate and cooperative with the home state in which they are registered. In many cases, as with EO and MPRI, much of SCs’ revenue originates from their home governments, who are long-term clients; therefore, SCs want to adhere to the policies and interests of their home states. For this reason, SCs usually share information with their home states regarding ongoing contracts even though such revelations are not required by domestic law. SCs’ directors and employees are often former members of the home states’ military or security institutions. These ex-military members rely on their personal contacts to secure contracts and the good graces of the government. They therefore retain allegiances for both personal and business reasons and will not challenge the authority or policies of these military and social institutions at home or abroad.

SCs limit their employers to legitimate governments and do not engage in ambiguous civil conflicts in which the population’s loyalty is in question, as in Zaire. Even if not restricted by the government licensing process, SCs restrict themselves in this manner so as to avoid the technical classification as mercenaries, which would result if they worked for insurgencies or rogue states, and to avoid intensified government and international scrutiny. EO has reiterated that its company is devoted to the advancement of legitimate African states and that it contracts with these governments “in order to do work designed to strengthen the self-determination of peoples, their internal stability and thus the possibility of putting economic development policies into practice.” In addition, because of fear for the loss of their personnel, SCs are unlikely to enter extremely violent conflicts that appear unwinnable. Involvement in controversial wars and loss of personnel have economic consequences (higher compensation for survivors and loss of expertise), morale implications (difficulty in future recruiting), and an impact on reputation (less likely to be hired by future customers).

SCs face an economic pressure to remain loyal to their employers. An employing government will not deliver the entire fee until the contract is fulfilled; therefore, an attack on the employer means that the SC will not be paid in full. Furthermore, the employer may represent long-term benefits by renewing its contract, signing new security deals, or requiring the services of an SC’s corporate allies, especially in the rebuilding of a country. Also, SCs do not want to be known in the international market as unreliable companies willing to turn on their employers at any point. For example, EO refused to negotiate with Mobutu Sese Seko of Zaire in part because of its contractual obligations to the MPLA government in Angola.

SCs have little incentive to commit wanton human rights violations. Since they are outsiders to the conflict, they tend to be apolitical and less passionate than those with personal stakes. Instead, SCs try to maintain a “hearts and minds policy” which makes their work in the field easier. Furthermore, contracting states are not likely to order SCs to perpetrate human rights abuses if the governments are trying to regain the support of the population and the respect of the international community. SCs employ elite military professionals who have worked together in the past, not ragtag miscreants. SC training
forces or combat units, then, are “well-drilled, disciplined force[s]” most often employed to professionalize a country’s forces. SCs’ ties to other corporate interests give them an economic stake in the peace and stability of a country and region. Businesses cannot grow amidst chaos, and lucrative rebuilding contracts cannot be signed while wars continue. This runs counter to the notion that mercenaries and SCs profit from the aggravation of war. Moreover the growing market for SCs is in the provision of quick, specialized training for government troops, not in combat engagement; therefore, many of the concerns over battlefield conduct relate only to those marginal SCs which deploy offensive troops.

In general, the market tempers and dictates SC behavior in such a way that most of the negative ramifications envisioned do not materialize. In addition to market forces, SCs are sensitive to international and domestic attention from the press, nongovernmental organizations, regional groups, their employers, and their home states. As a result of these market and political pressures, SCs have aligned their activities with the interests of their home states and have created a level of accountability.

C. Advantages

SCs fulfill an important role in international peace and security—a role which has been abdicated by states in many cases. SCs professionalize militaries and restructure officer corps, training them to function within democratic, civilian regimes and to fight under internationally accepted principles of engagement. Such training makes human rights abuses by native forces less likely. The training also diminishes the need for outside intervention in future conflicts since it makes the national forces self-sufficient. SCs allow small, struggling states to obtain quick and effective military expertise when other states are unwilling to devote their national forces or resources to aid besieged governments in their internal conflicts. Rather than being a threat to small states as mercenaries were in the past, SCs give small states a degree of independence from large state support or reliance on regional or international intervention. In part, this explains the attempt by the Papua New Guinea government to hire Sandline International to provide military training and guidance against the BRA. In addition, SCs give governments the military capability to bargain from a position of strength and bring an opposition movement to the negotiating table. The expense of SCs’ services forces rulers to make cost-benefit determinations as to the value of continuing to fight versus negotiating settlements.

Though SCs have economic stakes in the outcome of conflicts, they tend to be objective and apolitical since they are not personally entangled in the internal conflicts fueling the battles. Since most of the professional soldiers and trainers have other jobs and are paid well by SCs directly, they do not volunteer in order to maraud or plunder the countryside. In a situation of social chaos, as appeared in Sierra Leone, such dispassionate agents may be the only way of restoring order.

Low-scale civil conflicts represent a mixture of banditry and civil war fought for control of economic resources. They easily disintegrate into social chaos and spill over into neighboring countries, as seen in the Great Lakes region, West Africa, and the Sudan-
Uganda border region. SCs provide the stabilizing, objective internal force that the United Nations and the OAU (among other regional organizations) cannot offer. In addition, these companies protect humanitarian aid agencies and regional organizations, as seen with EO in Sierra Leone and with MPRI in the former Soviet Republics. EO’s effectiveness in providing stability and security has prompted the OAU to consider contracting with EO for continent-wide peacekeeping.

In sum, the SC market has developed because there is a need for such services in the world. SCs provide valuable services in restoring order and preventing internal conflicts from becoming international in scope in countries often ignored by the rest of the world. Yet, SCs have no role in countries with strong, consensual political and social institutions—one reason why EO’s presence in PNG was rejected. SCs professionalize militaries, promoting *jus in bello* principles and preserving the rights of noncombatants.

For exporting states, SCs are a means of aiding a regime or ally without committing national forces or compromising neutrality. Also, the SC gives ex-military officers, who could be a threat to their home states, a professional outlet for their services. SCs are an efficient tool for smoothing trouble spots in the international community, and can be vanguards for peace and stability.

**D. Conclusion**

Even though market forces currently seem to control many of the potential threats presented by the SC market, the market may evolve. The market may become saturated by SCs arising in the former Soviet Republics, other Western states, or Islamic states. This would force SCs to compete for controversial and risky contracts in more volatile environments. The home states of the new SCs may be less restrictive or may encourage SCs to work for rogue allies. SCs might then work for pariah regimes or brutal insurgencies. They could also resort to executing clandestine operations for third states. Similar results could be seen if legitimate employment opportunities diminish. However, given that at least thirty-four countries have contacted EO for its general training services, it does not seem that the legitimate market will shrink soon.

Regardless of what happens in the SC market, the potential problems with SCs are their lack of accountability and lack of transparency. The latter presents the greatest problem with EO, allowing speculation as to what its activities involve. SCs are important agents in a changing world order, and their potential excesses must be eliminated. However, total abolition of the services would simply drive organized mercenary activity underground and would result in the excesses hypothesized.

As a possible solution, SCs must continue to be tied to states. The best way of attaching SC activity to state responsibility is for each SC contract to be licensed by the exporting government. Several countries currently utilize export license regulations to monitor the defense service market, which includes SCs. Through domestic regulation and other state action, it must be made clear that SCs are just as much the responsibility of their home states as that of their contracting states. This may be the most effective way of controlling and monitoring SCs.
By allowing SCs to exist and not driving their services into a black market, states can regulate their activities and ensure that they are not used in illegitimate ways such as for challenging the sovereignty of states or to commit human rights violations. Regulation can isolate mercenaries employed by insurgencies and illegitimate regimes, the real threats to international peace and security. These rogue mercenaries are the unaccountable foreign soldiers that the international community has detested and tried to eradicate since the 1960s. Isolating these mercenaries would also serve to highlight the illegitimacy of the regimes or organizations employing them, as in the case of Mobutu’s regime in Zaire. In contrast, legitimate SCs could profit from regulation of the market. EO’s Eeben Barlow said he would welcome South African regulation of the industry: “There are too many cowboys and fly-by-nights, sometimes using our name, who are giving the industry a bad name.”

Regulating SCs, however, would remove the advantage of giving small states independence from large state support. To a certain extent, the proliferation of SCs would ameliorate this problem. Regulation would also eliminate the advantage to exporting states of appearing neutral in countries where its SCs operate. Creating a licensing regime for defense services would formalize the relationship that already exists between states and their SCs. SCs would become quasi-state agents, the actions of which home states would be responsible for.

VI. LICENSING PROCEDURES AS A MEANS OF REGULATING SCs

. . . There are other viable, yet less effective, options to control the SC market. The United Nations could attempt to formulate an international convention to regulate the market. Though there would probably be a general consensus among states as to the legitimacy of SCs, attempts to formulate an international convention would trigger innumerable disputes revolving around definitions, range of services, and inconsistent national laws. Even assuming that a convention could be formulated, the construction could take years, while the market continued to evolve without regulation.

Another avenue by which to control SCs is bilateral agreements between home states and contracting states. Generic treaties linked to aid packages could be signed between larger home nations and potential customer states. On the other hand, contract-specific treaties could explicitly delegate responsibility for SC activities and responsibilities. Countries could use these treaties to lock out SCs considered undesirable or dangerous for international security. Countries that see a potential need for SCs, however, are not likely to enter into limiting bilateral agreements unless induced by other advantages like increased military or economic aid from the exporting country. Such countries will desire to keep their options open so as to have the best selection of SCs; therefore, they are not likely to lock out specific SCs.

The most effective way of regulating the SC market is to assign responsibility to exporting states. Countries could pass legislation prohibiting particular types of activities or the export of defense services to certain regions or countries, as envisioned in the Diplock Report. The same definitional problems would persist under this approach. To best achieve state responsibility and SC transparency, states should regulate the SCs
through national registration and contract-specific licensing regimes. This would clarify SCs’ role as quasi-state agents, and states could take advantage of the efficiencies of the SCs in their foreign policy. Home states would be responsible for the excesses of their SCs and unregistered SCs could be sanctioned with civil and criminal penalties.

For example, the United States and Israel, among other countries, already apply to their SCs legislation adopted for arms and technology transfer. In the United States, the export of defense services is regulated under the provisions of 22 U.S.C. §§ 120-130, entitled “International Traffic in Arms Regulations.” Section 38 of the Arms Export Control Act authorizes the President to control the import and export of defense articles and defense services. Section 120.9 defines defense service as the following:

(1) The furnishing of assistance (including training) to foreign persons, whether in the United States or abroad in the design, development, engineering, manufacture, production, assembly, testing, repair, maintenance, modification, operation, demilitarization, destruction, processing or use of defense articles; or

(2) The furnishing to foreign persons of any technical data controlled under this subchapter (see § 120.10), whether in the United States or abroad.

Any company that either manufactures or exports defense articles or provides defense services must register with the State Department’s Office of Defense Trade Controls (DTC). Exporters must obtain a license pursuant to section 123 in order to conduct business abroad. The licensing procedure is handled by DTC, but Congress must be notified of certain types of exports, such as those of major defense equipment defined in section 120.8. In the course of approving a license application, DTC will obtain the advice of relevant agencies like the Department of Defense or country desks at the State Department. DTC resolves any disputes in accordance with larger U.S. policy goals. Firms may appeal adverse decisions by the DTC. The licenses are valid for four years or for the duration of the completion of the contract, whichever occurs first, and they may be amended as long as there are no material changes to the elements of the contract.

In addition to registration and licensing procedures, there is U.S. government oversight of SC activity at various stages. The “Blue Lantern Program” is a program that monitors the registration of companies, while the State Department maintains post-license, pre-shipment checks. The level of continued regulation depends on the nature of the contract and whether there are specific laws relevant to the contract. At all phases, the U.S. government promotes its interests by controlling U.S. exporters.

The Israeli government follows a similar procedure. In order for a national security company to work abroad, the company must have the approval of the Ministry of Defence. Export legislation controls military know-how, including “services relating to military equipment including information, instruction, training. . . military organization, operation, defense policy, [and] combat theory.” There is a two-stage licensing process for the export of controlled goods: the negotiation of the permits, and the determination of the appropriateness of export permits (licenses).
The foreign policy objective of Israeli defense export controls is “[t]o restrict the export of defense goods and technology where necessary to further the foreign policy of Israel, and to fulfill international obligations.” The Ministry of Defence will only approve licenses for export to legitimate governments and to governments whose activities are in accord with the foreign policy goals of Israel. Israel views the export of its military expertise as just as worthy of control as the export of military equipment. Israel clearly uses the export of its military expertise, including SCs, to further its interests abroad. The government ensures that SCs do not challenge or jeopardize its international relations by subjecting SCs to rigorous export controls tied directly to policy considerations.

This type of legislation reflects the reality that governments must face. In the international community, SCs represent the military expertise of their home state and implicate the presence of the home state regardless of how distanced the SC appears to be. The Mandela government has become aware that the activity of EO in Africa represents a “South African” presence. In addition, SCs are the repositories of valuable national information such as military strategy and “know-how.” Allowing such vital information to flow freely throughout the international market threatens the sovereignty of the exporting state. As a result, states realize that they must regulate and monitor SCs. Countries have used defense export controls and licensing procedures to harness the positive attributes of SCs while curbing the potential excesses.

The development of these laws and standards eventually will form an international norm in which states are expected to regulate the export of SC expertise. In such a regime, home states will begin to hold each other responsible for any excesses of their respective SCs. Although responsibility for SCs’ actions will be shifted to contracting states, it is clear, given current state practice, that home states realize the implications of SC contracts. SCs implicate their home states, and state responsibility flows from their activities abroad.

VII. CONCLUSION

In 1991, the military historian Martin van Creveld wrote a book entitled On Future War. He predicted that future conflicts are likely to take the form of small-scale wars waged by “groups we today call terrorists, guerrillas, bandits, and robbers, but who will undoubtedly hit on more formal ties to describe themselves.” He describes this transformation as follows:

‘The spread of sporadic small-scale war will cause regular armed forces themselves to change form, shrink in size, and wither away. As they do, much of the day-to-day burden of defending society against the threat of low-intensity conflict will be transferred to the booming security business; and indeed the time may come when the organizations that comprise that business will, like the condottieri of old, take over the state.

To some degree, van Creveld’s predictions are proving correct. In Africa, bands of armed profiteers wage low-intensity battles over valuable economic resources. Weak and disorganized states are ill-equipped to contain the chaos these bandits spread. Larger,
more capable states and the international community are usually unwilling to commit their national troops to help restore the internal order of collapsing states. Instead, weak states have found it more convenient, efficient, and reliable to hire foreign SCs which offer effective, objective, and immediate military expertise. The success and popularity of firms like EO and MPRI and the continued expansion of this market are a testament to the need for their services. Indeed, SCs are a modern-day permutation of *condottieri* built into the nation-state system.

However, SCs have not overtaken the state. The SCs’ reliance on their respective home states and contracting states for their economic survival have forced SCs to restrain their contracts. SCs work only for recognized regimes and act professionally in the field. They work in conjunction with international aid organizations and have not posed any threat to the sovereignty of states. On the contrary, SCs have reinforced the sanctity of the nation-state system by providing stability to faltering, yet legitimate regimes. They have fought the agents of chaos and have worked as restorers of the state. With their services, SCs quell the threats to the integrity of the international system and international peace and security which include undisciplined and untrained national troops, roving bandits, and terrorist groups. Ironically, SCs, as quasi-state agents, have become the nation-state system’s bulwark against destabilization, a role SCs have played effectively and diligently.

There is a debate in the international community about how to deal with SCs. The numerous U.N. resolutions and the International Convention regarding mercenaries were developed in response to the emergence of destabilizing mercenary forces in post-colonial Africa. The customary international law that has developed, therefore, reflects the international community’s concern with the unbridled violence of unaccountable individuals who could thwart movements of national liberation, launch coups, or aggravate ongoing civil unrest. The gravamen of the international community’s concern has been the accountability and transparency of these private actors. But these laws have not restrained the use of mercenaries:

Beyond formal resistance or the adoption of a stance of denying or minimizing the number of mercenaries and shared responsibility for their use, it is a fact that they are a resource used with a pragmatism that is morally and legally unacceptable because of what “mercenary” means and what a mercenary is worth as a professional of war and violence. Despite the condemnations contained in the resolutions of several United Nations bodies, *Governments whose power is illegitimate, armed insurgent groups and Powers acting through covert operations have been responsible for the existence of mercenary activities*, with a heavy toll on the peoples whose lives they affect.

The illegitimate employers of mercenaries—including terrorists, arms traffickers, drug cartels, alien governments, insurgencies—represents a union of nonstate actors that poses a direct threat to the state-based system. In this context, mercenary activities can only exist (as defined) within an illegitimate context. Foreigners hired by these types of organizations are mercenaries because their employers are illegitimate in the eyes of the state-based system.
Yet, in the case of SCs, there is a two-fold problem that has dovetailed into an international dilemma. First, SCs constitute a legal, corporate form of military service devoid of many of the negative consequences endemic in the deployment of “wild geese” mercenaries. Second, SCs have only worked for legitimate governments which are either democratically elected or are recognized by the United Nations. Therefore, although they in some ways resemble mercenaries, SCs operate professionally within the legitimate, state-based system. The regulation of mercenaries is limited to rogue military professionals who work for unaccountable bodies within the state system. Mercenaries, therefore, are distinguished from legitimate fighters or advisors by the illegitimacy of their employer.

Unlike the soldiers of fortune feared by all, SCs have promoted stability by empowering weak states against chaotic forces and have gained widespread public approval in the contracting countries. As repeat market players, SCs are constrained by profit motives and have accumulated an admirable human rights record. They have been accountable to their home states as a matter of good business, and their legal status has forced them to be fairly transparent.

However, there is no guarantee that the market forces that currently shape SC behavior in accordance with accepted norms of the international state system will not change. Companies like EO which are unconstrained by specific national regulations could become a destructive force and come to resemble an organized form of classical mercenarism. As a result, all states should regulate their SCs through licensing procedures that align SCs’ contracts with the policies of their home states. In this manner, SCs would become agents of the state and, therefore, accountable to the international community for their actions. Widespread adoption of these regulations would help create a new customary international law which would define SCs as contractors of the home and employing states.

Small states have taken it upon themselves to provide for their own security and not wait for the benevolence of larger states or the international community. In this sense, SCs have acted as a strike force for the United Nations in countries where the United Nations would like to intervene but has been unable to do so. The international community has already begun to consider directly taking advantage of the efficiencies these SCs provide. The United Nations and regional security organizations may consider using SCs as peacekeepers given the difficulty international organizations currently have of reacting quickly to escalating conflicts.

Several authors have proposed that the United Nations enlist former Gurkhas or the Foreign Legion to do its dirty work since these groups have a reputation of being efficient, effective, and objective in their services—traits often lacking in national troops sent as peacekeepers. There is widespread agreement that the United Nations must react more quickly to festering conflicts and that it is constrained by institutional and logistical shortcomings. As a result, the United Nations has been forced to sacrifice its objectivity by subcontracting some of its duties to regional hegemons in order to achieve its peacekeeping goals. In Haiti, Rwanda, Georgia, and Bosnia-Herzegovina, the United Nations has acquiesced to regional powers who have volunteered to lead peacekeeping
missions. Some U.N. policymakers have proposed the creation of a U.N. Legion, comprised of volunteers, which could react quickly and independently of state donations of national troops. SCs seem to fill a void in an international market thirsty for stability and security. Other international groups and NGOs like humanitarian aid organizations and missions in Sierra Leone and the former Soviet Republics may find it useful to contract with SCs to ensure that their aid is delivered properly and their workers are protected.

In a world where roving mobs can threaten the political integrity of a nation-state and threaten the tranquillity of the state’s neighbors, SCs provide a useful service that should be harnessed. In places like Uganda, where teenage thugs terrorize the northern farmlands and render the area uninhabitable, SCs may prove effective in restoring security and order.

Martin van Creveld stated the following about state security:

> The most important single demand that any political community must meet is the demand for protection. A community which cannot safeguard the lives of its members . . . is unlikely either to command their loyalty or to survive for very long. The opposite is also correct: [A]ny community able and, more importantly, willing to exert itself to protect its members will be able to call on those members’ loyalty even to the point where they are prepared to die for it.

SCs help states provide the security they need to retain order in an age of low-scale conflict. As a result, a new hybrid of the dog of war has evolved. What has emerged is a private organization that operates as a quasi-state agent to restore the sovereignty of the state, not to destroy it. It has been said that “[m]ercenaries are a product of a world which believes in the efficacy of force.” Similarly, SCs are a product of a world that needs security.


I. INTRODUCTION

Domestic law scholars and policymakers have long debated issues surrounding privatization. Over the past twenty years, the U.S. government has increasingly contracted with private organizations to perform a variety of functions—from health care, to education, to welfare, to prison management. While advocates of privatization have generally argued for the practice on efficiency grounds, critics have worried that, even if privatization may cut financial costs, it can threaten important public law values. Because many constitutional norms protect individuals only from government misconduct, and because courts have been largely unwilling to view such norms as applicable to private contractors, these critics have argued that privatization will dramatically reduce the scope of public law protections in the United States. Others have sought a middle ground,
arguing that privatization offers a means to extend public law values through the
government contracts themselves, in a process of “publicization.”

To date, however, none of these scholars has squarely confronted the growing
phenomenon of privatization in the international realm or its impact on the values
embodied in public international law. Yet, with both nation-states and international
organizations increasingly privatizing foreign affairs functions, privatization is now as
significant a phenomenon internationally as it is domestically. For example, states are
turning to private actors to perform core military, foreign aid, and diplomatic functions.
Military privatization entered the popular consciousness in early 2004, when private
contractors working as interrogators and translators for the U.S. government abused
detainees at Abu Ghraib prison in Iraq. But this kind of military privatization is only the
tip of the iceberg. Indeed, though the United States does not yet contract out direct
combat functions, we now frequently turn to private actors to provide logistical support to
those in combat on the battlefield as well as to aid in strategic planning and tactical
advice. Other states, such as Sierra Leone, have used private contractors to engage in
direct combat, and international organizations have weighed the possibility of using
private contractors to perform peacekeeping. In the foreign aid context, states and
international organizations are increasingly entering into agreements with private non-
profit and for-profit entities to deliver all forms of aid, including humanitarian relief,
development assistance, and post-conflict reconstruction. Even diplomatic tasks such as
peacekeeping negotiations are being undertaken by private actors in conjunction with
governments and international organizations.

All of this privatization in the international sphere raises the same sort of question for
international public law that domestic privatization raises for domestic public law: Will
privatization erode fundamental public law values, such as human rights norms, norms
against corruption and waste, and democratic process values? After all, international law
norms, like many domestic constitutional norms, traditionally apply only to states. The
Convention Against Torture, for example, generally prohibits only official misconduct.
Thus, we must ask: as more and more non-state contractors emerge on the international
scene, will these individuals and groups necessarily fall through the cracks of
international law and evade any public accountability?

My answer to that question is “no,” and in this Article I suggest that the domestic U.S.
administrative law literature may provide a useful set of responses to privatization that
has been largely overlooked by international law scholars, policy-makers, and activists.
In particular, I argue that possibilities for extending public law values inhere in the
privatized relationship itself, particularly in the government contracts that are the very
engine of privatization. Indeed, the contracts governments enter into with non-state actors
can include many provisions that would help to create both standards of behavior,
performance benchmarks, and a means of providing some measure of public
accountability. While such contractual provisions are not a panacea, they may be at least
as effective as the relatively weak enforcement regime of public international law. At the
same time, by considering international privatization, I seek to open what I believe could
be a fruitful dialogue between domestic administrative law scholars and international law
scholars about possible responses.
Significantly, while domestic scholars of privatization have not yet turned their attention to foreign affairs privatization, international law scholars have not really focused on privatization at all, and, in any event, have not seriously considered contract as a source of solutions to the potential threat to public law values that privatization may seem to pose. Of course, international law scholars have recognized concerns about how to apply international legal norms to non-state actors in general. But “non-state actors” is too broad a category because a private contractor is very different from, say, a guerrilla soldier. In particular, because privatization involves an increasing contractual relationship between governments (or international organizations) and private actors, contractual mechanisms for importing public accountability are potentially available with regard to privatization, whereas they obviously are not relevant to many other instances in which non-state actors play a role in the international sphere.

Moreover, to the extent that international law scholars and policy-makers have proposed any solutions to potential problems created by privatization, these proposals fall far short. At the extreme, some have argued that the best response to military outsourcing, for example, is simply to oppose it altogether, because military functions are somehow “inherently” governmental or because state bureaucracies can be better monitored and held to account in court than private contractors can. However, those who simply resist privatization are misguided because the trend toward outsourcing of foreign affairs functions previously performed by state bureaucracies (at least in the recent past) is probably irreversible. The privatization train has not only already left the station, but has gone far down the track. Indeed, even those who seek to send the train back home should favor alternative solutions in the interim, because any return is likely to take a very long time.

Others have argued that private actors with significant impact in the international sphere should be more formally brought within the normative framework of international law. Thus, with each wave of non-state actors—such as guerrilla movements, terrorists, non-governmental organizations, and corporations—many international law practitioners and scholars have considered expanding the coverage of public international law to apply to each group. They have therefore contended either that states should (by treaty or customary international law) develop new norms that apply directly to these categories of non-state actors, or that any “state action” requirements contained in existing norms (again either in treaties or customary international law) should be interpreted expansively to apply to non-state actors linked to the state. At the same time, these scholars and practitioners have tended to focus on the need for courts and tribunals—in many cases new ones—to apply and interpret these norms.

Yet this approach, though it is important because it results in the articulation of norms in the international sphere, can have only a limited effect. Even if the proposed courts and tribunals are established and fully functioning, and even if they expand the norms of international law to apply to the broad range of privatized action, these tribunals will never have the capacity to hold more than a limited number of individuals (and groups) to account. Accordingly, we need to seek alternative mechanisms for extending and implementing public law values to privatized actors in the international sphere. And though literature on corporate responsibility, NGOs, soft law, and transnational networks
has attempted to address some informal modes of accountability, international law scholars have so far not sufficiently discussed the possibility of contract.

This Article begins by laying out the scope of the problem, surveying the extent of foreign affairs privatization, the potential threat it poses to public law values, and the failure of current outsourcing contracts to address this problem. Using Iraq as a case study, I examine the publicly available military contracts as well as contracts to provide foreign aid, and I suggest serious deficiencies in the contracts thus far. Then, drawing on examples and insights from the domestic privatization literature, I set forth nine ways in which contractual provisions could be used to extend and enforce public law values in the foreign affairs privatization context. Specifically, I suggest that contracts be drafted to: explicitly extend relevant norms of public international law to private contractors, delineate training requirements, provide for enhanced monitoring both within the government and by independent third-party monitors, establish clear performance benchmarks, require accreditation, mandate self-evaluation by the contractors, provide for governmental takeovers of failing contracts, include opportunities for public participation in the contract negotiation process, and enhance whistleblower protections and rights of third-party beneficiaries to enforce contractual terms. And, because these public values would be embodied in that quintessential private law instrument—the contract—they would more readily come within the purview of domestic courts or private arbitral bodies and so would rely less on international public law enforcement mechanisms (though those are possible as well). As a result, these contractual provisions may at least make some progress in attempting to ensure that private contractors are accountable both to the publics they serve and to those who are most affected by their work.

Of course, one might think that these proposals are unrealistic because one of the main reasons governments privatize is precisely to avoid the kind of accountability I propose. Yet governments are not monolithic, and there are undoubtedly many people within bureaucracies, such as contract monitors, who honestly wish to do their job and would therefore welcome (and lobby for) contractual mechanisms that increase accountability. In addition, non-governmental organizations (NGOs) and international organizations can sometimes pressure states to adopt oversight regimes such as the ones I discuss. The problem is that the policymakers and scholars have not sufficiently focused on privatization or the possible accountability mechanisms that could be embodied in contracts. Thus, this Article seeks both to raise awareness about the ways in which contractual provisions might embody public law values and to stimulate a broader-ranging debate about the best way to respond to privatization in the international context.

II. FOREIGN AFFAIRS PRIVATIZATION AND THE THREAT TO PUBLIC LAW VALUES

B. The Threat to Public Law Values

. . . Just as the protections contained in the U.S. Constitution are generally viewed as prohibitions on state misconduct only, the principal international human rights and humanitarian law instruments of the twentieth century—the United Nations Declaration on Human Rights, the International Covenant on Civil and Political Rights, the
International Covenant on Economic, Social and Cultural Rights, the Genocide Convention, the Convention Against Torture, the Fourth Geneva Conventions, and the Rome Statute of the International Criminal Court—were drafted primarily with states in mind. As such, at least in the conventional account of public international law, states are seen as both the primary parties to the treaties and the central bearers of rights and responsibilities. These instruments do grant individuals rights, of course—such as the right to be free from torture, cruel, inhuman or degrading treatment, or the right to a fair trial—but these are generally conceived primarily as rights against the state. Conversely, individuals can be held criminally liable, but usually only if some connection to the state is demonstrated. And while there are some exceptions within the overarching framework of public international law—for example, individuals can be convicted for genocide and crimes against humanity (as defined in the recent statute of the International Criminal Court) regardless of any connection to a state apparatus—state action (or at least a link to it) still remains at the core of most conceptions of international law liability.

The private contractor interrogators and translators implicated in the abuse at Abu Ghraib prison provide a notable example of how these non-state actors might fall through the cracks of this traditional, state-centered approach to public international law. Although the Geneva Conventions and the Convention Against Torture clearly prohibit any abuses committed by U.S. military personnel, the treaties’ applicability to non-state actors is ambiguous, and fora for holding non-state actors accountable are limited. The U.S. government’s use of private contractors to transport terrorism suspects to countries known to practice torture has raised similar questions because again, while the Convention Against Torture prohibits governments from taking such actions, its applicability to private actors is ambiguous.

Private military companies engaging in direct combat also arguably fall through the cracks of current international law provisions, despite probably being the most notorious for committing atrocities. Although multiple treaties ban the use of certain categories of mercenaries outright, broad gaps in the definition of “mercenary” leave most types of work by private military companies outside the treaties’ prohibitions. For example, in Sierra Leone in the 1990s officers of Executive Outcomes, working under contract with the government, reportedly ordered employees carrying out air strikes against rebels to “[k]ill everybody,” even though the employees had told their superiors they could not distinguish between civilians and rebels. While such a command would almost certainly constitute a war crime if ordered by a military or civilian authority in the chain of command, it is less clear that such actions committed by private contractors would qualify, at least absent inquiry into the extent of the contractor’s link to the government.

Abuses committed by private actors who deliver aid also raise complicated questions about the application of international law. Although aid workers do not by any means regularly mistreat aid beneficiaries, such incidents occur more often than one might suspect. For example, employees of DynCorp Inc., a private corporation that was charged with training police in Bosnia in the 1990s under a contract with the U.S. government, were “implicated in a sex-trafficking scandal” involving acts of rape, sexual abuse, and exploitation. Even staff members of not-for-profit organizations have at times been implicated in abuses. Indeed, a recent study of refugees and internally displaced persons

842
Draft — do not distribute!
in West African camps in Guinea, Liberia, and Sierra Leone reported widespread rape and sexual exploitation of women and children by many actors, including aid workers. The aid workers and peacekeeping forces allegedly relied on their positions of relative power to use “the very humanitarian aid and services intended to benefit the refugee population as a tool of exploitation.” In some camps, it appears that even necessities such as using a toilet were sometimes conditioned on the willingness to perform sexual favors. Although, as in the military context, such abuses committed by governmental actors generally violate international agreements, the same acts committed by non-state actors fall into a gray area.

Even outside the human rights context, the principal regional treaties seeking to deter corruption, for example, apply primarily to misconduct involving governmental actors. Yet, foreign aid contractors have been implicated in fraud and waste. Indeed, Kellogg Brown & Root’s more than $10 billion in contracts with the U.S. government in Iraq “have been dogged by charges of preferential treatment, overbilling, cost overruns, and waste.” Elsewhere, employees of Custer Battles, a company that was awarded two $16 million contracts by USAID to provide security for the Baghdad airport and distribute Iraqi dinars, reportedly chartered a flight to Beirut with $10 million in new Iraqi dinars in their luggage—which were promptly confiscated by Lebanese officials. The company also set up sham Cayman Islands subsidiaries to submit invoices, and regularly overcharged for materials—in one case charging the United States $10 million for materials that it purchased for $3.5 million. In short, corruption and fraud have been rampant in the Iraqi contracts. Yet, legal oversight (and democratic accountability) is limited because such contractors operate beyond many of the transparency rules that would apply to government entities.

Thus, widespread privatization potentially threatens a wide variety of public law values. One response to this problem, of course, is to interpret (or amend) the international law norms themselves either to remove any state action requirement, or at any rate to construe such a requirement leniently. Indeed, at least some of the conventional state-centered story of international law that I have recounted has long been subject to challenge. For example, the U.N.’s Draft Articles on Responsibility of States for Internationally Wrongful Acts aims to make clear that the “conduct of any State organ shall be considered an act of that State under international law,” and that a person’s conduct shall be attributed to the state if he or she is acting on the state’s instructions or under the state’s direction. Likewise, courts and tribunals have at times applied principles of state responsibility for instrumentalities to impute the liability of companies onto states. And some courts have suggested that certain international law norms, such as the laws of war, can be used to hold non-state actors directly accountable, at least in some circumstances. Alternatively, non-state actors can be held liable under theories of complicity or aiding and abetting. Thus, it would be wrong to characterize international law as completely impotent with regard to private contractors.

Yet, though I am sympathetic to efforts to revise or interpret the norms of public international law to apply to private contractors, I argue that such efforts should not be the only response to privatization in the international realm. Indeed, public international law norms are imperfectly enforced in the best of circumstances, and any interpretational
ambiguities with regard to contractors only compounds the practical difficulties. Thus, those concerned that public values may be lost in a privatized world would be well-advised to look in other directions as well. And, as we will see, the contractual relationship that creates the very structure of privatization may itself offer a means of promoting and enforcing public law values, and it is to such avenues of accountability that we now turn.

III. CONTRACT AS A TOOL TO EXTEND AND ENFORCE PUBLIC LAW VALUES

Contracts between governmental entities and the private organizations providing services can themselves serve as vehicles to promote public law values. Contractual terms can specify norms and structure the contractual relationship in ways that spur contractors to implement those norms. Thus, although typically conceived as the quintessential private law form, contracts used in this way can be a tool to “publicize” the privatization relationship.

Administrative law scholars have recently explored this insight in the domestic context. Most notably, Jody Freeman has suggested that states “could require compliance with both procedural and substantive standards that might otherwise be inapplicable or unenforceable against private providers.” Yet this work has focused on privatization of healthcare, welfare, and prisons in the United States and has not considered privatization of military, foreign aid, and diplomatic activities. Meanwhile, as noted previously, few if any international law scholars, policymakers, or NGOs have considered the possibilities of using contractual terms in the international context. Accordingly, this Part seeks to bridge the gap between domestic administrative law and international law scholarship by exploring a variety of contractual mechanisms that might be used to extend public law values to privatized foreign affairs.

Specifically, I discuss nine contracting practices that could be deployed in the foreign affairs arena: (1) incorporating public law standards in contractual terms; (2) requiring that private contractors receive training; (3) enhancing contractual monitoring, both by internal governmental actors and third parties; (4) requiring that contractors receive accreditation from independent organizations; (5) laying out clear performance benchmarks; (6) mandating contractor self-evaluation; (7) enhancing governmental termination provisions and allowing for partial governmental takeover of contracts; (8) allowing for beneficiary participation in contract design; and (9) strengthening enforcement mechanisms, including greater whistleblower protections and more opportunities for third-party beneficiary suits. In considering these possibilities, I will use Iraq as a case study, examining all of the publicly available contracts the U.S. government has negotiated to support the U.S. military or to provide for foreign aid to Iraq. Nevertheless, these same principles would apply to other types of contracts negotiated by states or international organizations with contractors providing a variety of foreign affairs functions.

The contractual mechanisms I discuss are particularly important in the foreign affairs context because many of these contracts are negotiated in secret, without competition, on a “no bid basis,” based on exceptions to the normal requirements of the Federal

Draft — do not distribute!
Acquisition Regulation (FAR). For example, with respect to the U.S. government’s foreign affairs contracts in Iraq, in many cases it is impossible for the public or a watchdog group even to obtain the text of the contracts, either because government officials have kept them secret for security reasons, or because the contractors have exercised what is essentially a veto, under the Freedom of Information Act (FOIA), for certain types of commercial information. Problems posed by secrecy are reinforced by problems of conflict of interest because many of the contracts are awarded to firms run by former government personnel. A 2003 study by the Center for Public Integrity reports that sixty percent of the companies that received contracts in Iraq or Afghanistan “had employees or board members who either served in or had close ties to the executive branch for Republican and Democratic administrations, for members of Congress of both parties, or at the highest levels of the military.” Thus, it is essential that, at the very least, the contracts themselves incorporate public values.

A. Incorporating Public Law Standards in Contractual Terms

First, of course, the contracts could explicitly require that the contractors obey the norms that implement public law values. Specifically, the terms of each agreement could provide that private contractors must abide by relevant legal rules applicable to governmental actors. Such contractual terms would obviate the need to show that the private actors were functioning as an extension of government so as to satisfy any state action requirement that might arise under domestic and international legal regimes. Instead, the norms applicable to governmental actors would simply be part of the contractual terms, enforceable like any other provisions, regardless of state action.

In the domestic setting, such provisions are commonplace. As a term in their contracts with privately run prisons, for example, many states require compliance with constitutional, federal, state, and private standards for prison operation and inmates’ rights. In addition, contractual agreements may require contractors to provide for hearings and review of contractor actions.

The U.S. government’s military and foreign aid contracts in Iraq, by contrast, are woefully inadequate on this score. To be sure, a 2005 Department of Defense (DOD) document providing general instructions regarding contracting practices does state that contractors “shall abide by applicable laws, regulations, DoD policy, and international agreements.” Yet, of the sixty publicly available Iraq contracts, none contains specific provisions requiring contractors to obey human rights, anticorruption, or transparency norms. The agreements between the U.S. government and CACI to supply military interrogators starkly illustrate this point. The intelligence personnel were hired pursuant to a standing “blanket purchase agreement” between the Department of the Interior and CACI, negotiated in 2000. Under such an agreement the procuring agency need not request specific services at the time the agreement is made but rather may enter task orders as the need arises. In 2003, eleven task orders, worth $66.2 million were entered (none of which was the result of competitive bidding). The orders specify only that CACI would provide interrogation support and analysis work for the U.S. Army in Iraq, including “debriefing of personnel, intelligence report writing, and screening/interrogation of detainees at established holding areas.” Significantly, the orders do not
expressly require that the private contractor interrogators comply with international human rights or humanitarian law rules such as those contained in the Torture Convention or the Geneva Conventions. Likewise, although the contractors are subject to international and domestic laws prohibiting the bribery of government officials, as well as the general terms of the FAR, none of the contracts specifically prohibits the contractors themselves from accepting bribes, an area that remains ambiguous in domestic and international law. Similarly, the contracts do not provide terms specifying the applicability of FOIA, which would help make contractor activities more transparent.

**B. Requiring That Private Contractors Receive Training**

Foreign affairs contracts could also explicitly require that contractors receive training in activities that would promote public law values. Such training, as a contractual requirement, could help instill in contractor employees a sense of the importance of these values. At the same time, training could provide employees with concrete recommendations about how to implement these values in specific, challenging situations.

Again, in the domestic setting such training provisions are commonplace. A standard term in state agreements with companies that manage private prisons, for example, requires companies to certify that the training they provide to personnel is comparable to that offered to state employees. Such training would normally include instruction concerning legal limits on the use of force and examples of what those limits mean in circumstances likely to arise in the prison setting.

Yet, while the 2005 DOD instructions require documentation of training concerning appropriate use of force, none of the publicly available Iraq contracts appears to require such training. Indeed, although a few of the agreements require that contractors hire employees with a certain number of years’ experience, none specifies that the contractor must provide any particular training at all. For example, the U.S. government’s agreement with Chugach McKinley, Inc. to screen and hire a broad range of military support personnel—from doctors to “special mission advisers”—says nothing about whether such personnel will receive training in applicable international law standards, even though such personnel may be in a position to commit abuses. The U.S. government’s agreements with CACI to provide interrogators are likewise completely silent on whether interrogators will receive education in international humanitarian and human rights law, training that U.S. military interrogators would normally receive. Not surprisingly then, an Army Inspector General report on the conditions that led to the Abu Ghraib scandal concluded that 35% of CACI’s Iraqi interrogators did not even have any “formal training in military interrogation policies and techniques,” let alone training in international law norms. This omission is particularly glaring given the highly volatile Iraqi environment.

Anti-corruption training would also be useful for foreign affairs contractors generally, and for contracts in Iraq specifically. Iraq ranks among the worst countries in the world on Transparency International’s corruption index, and it is no surprise that such corruption reaches U.S. contractors operating there. Indeed, one former Coalition
Provisional Authority (CPA) official, Alan Grayson, has asserted that lack of employee screening and training led to the shocking abuses committed by Custer Battles. Yet such contracts say nothing about training for contractors in practices to avoid corruption. And while training requirements undoubtedly would increase the cost of the contracts, the fraud and waste that could be deterred with better training might well offset such increases.

C. Enhancing Contractual Monitoring, Both by Internal Governmental Actors and by Third Parties

Provisions could also be made for increased contract monitoring, which could provide an important check on abuses. Such monitoring should include, to begin with, sufficient numbers of trained and experienced governmental contract monitors. At the same time, governmental ombudspersons—leaders of independent offices charged with providing enhanced oversight—serve as an important supplement to the contract monitors. Thus, at a minimum, it is essential that government agencies devote enough resources to ensure that these requirements are implemented in a meaningful way. In addition, outside independent non-governmental organizations, both for-profit and non-profit, can serve an important function by monitoring contracts.

Contracts for services in the domestic context regularly include this three-tiered monitoring structure: government personnel assigned as contract monitors, supplemented by agency actors such as ombudspersons, further supplemented by independent outside groups. In the privatized health care context, for example, where private nursing homes receive Medicaid funding and private hospitals receive Medicare and Medicaid support, the trend is toward agreements that require a state-appointed contract manager. Federal agencies such as the Department of Health and Human Services (whose Inspector General issues reports on contracts with private hospitals that receive public funding) and the Health Care Financing Administration (which exerts fairly tight control over private nursing homes receiving Medicaid funding) also have significant oversight authority. In addition, third-party independent organizations play an important role. For example, the Joint Commission on Health Care and Accreditation of Health Organizations (JCAHO), a private organization of professional associations, certifies health care institutions for compliance with federal regulations and state licensure laws.

Foreign affairs contracts currently provide for far less monitoring. To be sure, the statutory and regulatory scheme includes provisions for governmental contract monitors, supplemented by inspectors general of the respective agencies responsible for the contracts, as well as for auditing of contracts by independent private accounting firms. Yet, the work of these monitors focuses primarily on whether the contractors are keeping adequate accounts and refraining from fraud and bribery. Contracts say little about human rights norms, and governmental contract monitors and ombudspersons are not ordinarily focused on these values when scrutinizing contractors. To the extent that independent third-party groups are empowered to monitor under the contract, they tend to be auditing firms, whose expertise lies in financial matters, not in international human rights or humanitarian law. Foreign affairs contracts rarely, if ever, provide for monitoring by independent groups with expertise in this area.
Moreover, in practice, foreign affairs contracts tend to escape even this limited oversight. This is because many of the monitoring requirements tend not to apply in emergency situations, which are, of course, precisely the occasions when military intervention or humanitarian relief efforts and post-reconstruction aid are most likely. Thus, ordinary contracting procedures, such as competitive bidding, are often waived. In addition, many of the contracts are written as cost reimbursement contracts, often termed “cost-plus” agreements, under which the government reimburses the contractor for costs incurred in providing a service, plus a fee that is calculated as a percentage of the cost. Though often criticized as leading to waste and abuse, such contracts become the norm in emergency situations, rather than the exception. At the same time, too few contract monitors are appointed, those who are appointed lack expertise, and ombudspersons are not given the resources they need to do an effective job.

The monitoring of the Iraq contracts, or virtual lack thereof, provides a salient example. The government agencies with responsibility for the contracts—primarily USAID, the DOD, and the now dismantled Coalition Provisional Authority—devoted extraordinarily minimal resources to monitoring. For example, USAID has responsibility for approximately $3 billion in reconstruction projects, but the agency had only four contract monitoring personnel on the ground as of March 2003. In fact, due to the difficulties of monitoring contracts with so little staff, USAID determined to contract out the monitoring function itself! Likewise, a recent DOD Inspector General study concluded that more than half of the Iraq contracts had not been adequately monitored. This fact is not surprising given that DOD’s acquisition workforce was reduced by more than half between 1990 and 2001, while the department’s contracting workload increased by more than twelve percent. In addition, those who were assigned to monitor contract performance were often inadequately trained. Finally, in an ironic twist, private contractors themselves are often hired to write the procedural rules governing contracting rules and monitoring protocols, thus leading to further conflict-of-interest problems. Indeed, the DOD handbook on the contracting process was drafted by one of its principal military contractors.

The CPA was plagued with similar problems. A recent report notes that the CPA hadn’t kept accounts for the hundreds of millions of dollars of cash in its vault, had awarded contracts worth billions of dollars to American firms without tender, and had no idea what was happening to the money from the Development Fund for Iraq (DFI) which was being spent by the interim Iraqi government ministries.

One former CPA official has observed that, as a result of poor oversight, “contracts were made that were mistakes, and were poorly, if at all, supervised [and] money was spent that could have been saved, if we simply had the right numbers of people.” For example, even devoting a single staff person to the two $16 million Custer Battles contracts that gave rise to multiple instances of fraud and abuse would have saved at least $4 million.

Finally, the dispersal of authority to issue foreign affairs contracts across multiple agencies creates interagency communication problems and conflicts of interest that impede oversight. For example, officials at different agencies use different methods to calculate the costs of contracts, and these methods may also vary from those used by the
companies themselves. In addition, because agencies can earn fees for facilitating other agencies’ contracts but are not adequately held to account for monitoring those contracts, agencies have incentives to sponsor other agencies’ contracts but little incentive to supervise them. These arrangements can lead to abuse, as occurred in the case of the Department of the Interior sponsorship of DOD’s task orders for intelligence services at Abu Ghraib prison under an existing contract between CACI and the Interior Department.

In short, the foreign affairs contracts could provide far better protections for public law values through greater monitoring. Although the statutory and regulatory regime contemplates a combination of supervision by contract monitors, independent agency oversight through inspectors general, and limited financial auditing by third-party entities, these provisions have not worked well in practice due to insufficient staffing and resources, combined with the large number of contracts. To be sure, statutory and regulatory reforms could address these problems. But, alternatively, the contracts themselves could remedy these deficiencies to some extent, by specifying greater numbers of monitors and requiring that they possess a certain degree of training, as well as by allowing for independent oversight by third-party groups such as the International Committee of the Red Cross (ICRC).

D. Laying Out Clear Performance Benchmarks

Of course, to some degree increased contract monitoring can only be effective to the extent that the contracts have clear benchmarks against which to measure compliance. In the domestic context, commentators and policymakers have long urged that contracts include benchmarks, and rigorous performance standards regularly appear in contracts. Scholars have argued that, ideally, performance-based contracts should “clearly spell out the desired end result” but leave the choice of method to the contractor, who should have “as much freedom as possible in figuring out how to best meet government’s performance objective.”

These ideas have been implemented most notably in contracts with private prisons. For example, under the model contract for private prison management drafted by the Oklahoma Department of Corrections, contractors must meet such delineated standards for security, meals, and education. They must also certify that the training provided to personnel is comparable to that offered to state employees. In Texas, contractors must abide by similar terms and, in addition, must “establish performance measures for rehabilitative programs.” In addition, the American Correctional Association is revising its accreditation standards to include performance measures, and the Office of Juvenile Justice and Delinquency Prevention is developing performance-based standards for juvenile correctional facilities. Commentators have noted, further, that performance measures for private prison operators could include both process measures such as the number of educational or vocational programs, or outcome measures such as the Logan quality of confinement index, the number of assaults, or the recidivism rate. . . . Because no single statistic adequately captures “quality,” and because focusing on any single measure could have perverse
effects, performance-based contracts should tie compensation to a large and rich set of variables.

Privatized welfare programs have also experimented with performance measures as a means to improve quality. In 1996, Congress authorized the implementation of welfare programs “through contracts with charitable, religious, or private organizations.” Since then, states have increasingly contracted with such organizations, and many of these contracts contain performance benchmarks and output requirements. For example, under a performance-based system, a welfare contractor might receive financial rewards for increasing the percentage of program participants who receive job placements.

The foreign affairs contracts are notably less rigorous in providing for performance measures. Although military service contracts are difficult to evaluate because so many of them are not publicly available, contract officers familiar with the contracts have remarked on their generally vague terms. And the fact that they are often indefinite delivery/indefinite quantity (ID/IQ) contracts adds to their open-ended quality. Under this structure, the government awards a contract that does not specify how many services or goods will be necessary or the dates upon which they will be required. These additional details are specified in subsequent task orders, which themselves are often vague because the task orders need not pass though the same degree of supervision as the initial contract award. Of course, such contracts may sometimes be necessary, because the government cannot know in advance precisely what will be required or for how long. Yet the lack of any administrable standards in these contracts can lead to significant abuses.

Of the publicly available Iraq contracts for military services, it is striking that none contains clear benchmarks or output requirements. Instead, they are phrased in amorphous language that provides little opportunity for compliance evaluation. For example, a contract between the U.S. government and MPRI to provide translators for government personnel, including interrogators, simply provides that the contractors will supply interpreters. The agreement says nothing about whether the interpreters must be effective or how effectiveness might be measured. Similarly, the CACI task orders for interrogators specify only that CACI will provide interrogation support and analysis work for the U.S. Army in Iraq, including “debriefing of personnel, intelligence report writing, and screening/interrogation of detainees at established holding areas.” Other than these broad goals, the task orders say little more. To be sure, security concerns may require some degree of vagueness. Nonetheless, the task orders could be much more specific about training requirements, standards of conduct, supervision, and performance parameters.

Turning to the foreign aid context, agencies tend to promote the use of results-based agreements, under which contractors must demonstrate specific, tangible results that are to be evaluated by the agency. Yet in practice many such agreements do not actually contain any results-based requirements, often because the aid, particularly in emergency relief settings, is provided on an expedited basis to organizations with very small staffs. With regard to Iraq, for example, a review of the publicly available USAID agreements reveals that only a few set forth specific performance benchmarks or requirements.
To be sure, performance benchmarks that are too strict can pose problems. As scholars of domestic privatization have noted, discretion can serve useful goals; indeed, discretion is in part what makes privatization desirable, as private contractors have more flexibility than rulebound bureaucratic actors to pursue innovative approaches. Output requirements that preserve flexibility about the means to achieve those results are therefore the most effective. But even carefully tailored output requirements can go awry, as when, for example, private welfare providers “cream” those accepted into their programs in order to increase the percentage of those who receive job placements. Moreover, output requirements can sometimes give contractors tunnel vision, leading them to focus only on the benchmarks, thereby missing opportunities to achieve wider benefits. A recent study of the enhanced “auditing” that accompanied privatization in Thatcherite Britain, for example, suggests that narrow output requirements steered organizations and individuals away from broader, more diffuse, social goals.

In addition, by their very nature, results-based contracts raise difficult questions about how best to measure output. Creating benchmarks may be relatively straightforward if the project at issue involves simply building a bridge or dam, but it is very difficult to measure intangibles, such as fostering human development or building civil society. Likewise, short-term results, such as whether food aid was delivered, are much easier to measure than longer-term systemic efforts to alleviate poverty, provide education, and so on. As a consequence, results-based contracts tend to put more emphasis on short-term delivery of services rather than longer-term impact. Finally, contractual output requirements do not, of course, necessarily ensure compliance because contractors may simply fail to meet their goals. In addition, even the most detailed performance requirements and standards inevitably leave considerable discretion to the contractor.

Nonetheless, despite problems with overly rigid performance benchmarks, the foreign affairs contracts (at least those that are publicly available) appear to fall at the opposite end of the spectrum. Indeed, they possess so few benchmarks and output requirements that they contain no meaningful evaluative criteria whatsoever. In such circumstances, enhanced performance benchmarks could be a useful contractual tool.

E. Requiring That Contractors Receive Accreditation from Independent Organizations

Another contract-based tool for promoting public law values is accreditation. Independent organizations, often consisting of experts or professionals in the field, can evaluate and rate private contractors. These ratings can then be used in the contracting process because agreements can require that contractors receive certain rankings. Or, governmental entities or international institutions, such as the United Nations, could develop accreditation regimes.

Again, the domestic context offers a particularly rich set of examples that could provide useful lessons in the foreign affairs setting. For example, in the field of publicly funded, privately provided health care, JCAHO accredits hospitals receiving Medicare and Medicaid funding. Indeed, such accreditation is required by statute as well as by contract. State laws or contractual terms also often specify that health maintenance organizations must receive accreditation by the National Committee for Quality Assurance (NCQA), an
independent non-profit organization, before receiving public funding. Until recently, NCQA certification was primarily voluntary, offering health maintenance organizations an advantage when competing for lucrative health care delivery contracts. When states became managed care purchasers, however, they adopted NCQA as a benchmark of quality.

Similarly, many contracts with private prison operators require companies to receive accreditation by the American Correctional Association (ACA). An organization of correctional professionals that has existed for over a century, the ACA accredits prisons and provides training for prison personnel while also setting standards that apply to virtually every aspect of prison operation. Not only has ACA accreditation become a standard contract requirement, but federal courts have used ACA standards to interpret constitutional and statutory provisions. Even private investors look to accreditation as an indication of quality. Thus, the accreditation requirement creates significant compliance incentives.

Privatized education regimes such as charter schools have also considered accreditation by independent organizations as a means of ensuring quality. The focus of many independent organizations on facilities and administrative processes over underlying educational quality has led some critics to charge that educational accreditation is relatively ineffective. Nonetheless, commentators have advocated improved accreditation procedures and greater use of such accreditation to promote public law values.

Indeed, domestic administrative law scholars have noted that these independent, private accrediting entities are effectively setting the standards that give meaning to public law values. In that regard, the relative insularity of the standard-setting and accreditation process may undermine the ability of broader groups, including consumers and the public at large, to participate in the process. There is also the concern that private accreditors in some cases might be too close to the contractors, and therefore too lenient. Nevertheless, even critics agree that the standards are often much better than those that would be developed by agency bureaucrats, and despite the imperfections, accreditation has served as an important check on the contracting process.

In contrast, accreditation is glaringly absent in the foreign affairs context. Human rights organizations, governments, and the United Nations have begun to encourage corporations, particularly those in the extraction industries, to comply with voluntary labor, environmental, and human rights standards. A consortium of NGOs that deliver humanitarian relief have initiated the SPHERE project, which is an effort to set standards for the provision of humanitarian aid, including specific guidelines for field operations, training, and self evaluation. And an industry-founded association of private security companies, the International Peace Officers Association (IPOA), has begun to construct a comprehensive code of conduct that includes human rights standards. Nevertheless, neither the U.N., nor domestic governments, nor outside groups concerned with potential abuses by foreign affairs contractors have so far undertaken serious efforts either to harness these nascent accreditation initiatives or to promote other accreditation projects.
This failure is particularly striking in the Iraq context. Not one of the available contracts for aid or military services requires that the entities receiving the contracts be vetted or accredited by independent organizations. For example, unlike domestic prison contracts, which routinely require accreditation by ACA and compliance with a comprehensive set of standards, the contracts with CACI to provide interrogators at Abu Ghraib contain only the most basic guidelines and make no mention of human rights compliance or accreditation requirements. The contract between the U.S. government and Dyncorp to provide law enforcement advisers to train Iraqi police similarly contains no provision mandating that Dyncorp be accredited, even though Dyncorp employees were implicated in sex abuse when performing under a similar contract in Bosnia. Likewise, although contracts could require that humanitarian aid organizations agree to the SPHERE guidelines in order to receive contracts, no such requirement has been imposed.

Yet, such accreditation would seem to be particularly important in the foreign affairs area, where, as discussed previously, security concerns and special considerations often eliminate competition in the contracting process, resulting in contracts that are structured without the usual market controls. Significantly, the problem is not only that international organizations and domestic governments neglect to require accreditation in their contracts, but also that NGOs and other independent groups have not sought a robust accreditation role. After all, more NGOs could, like the SPHERE Project’s efforts in humanitarian aid, begin to rate military contractors independently, regardless of whether the government contracts require such accreditation. These ratings might then become an industry standard that the government could be persuaded to use as a contracting factor. This is what occurred with NCQA in the domestic health care context. And, even if agency officials negotiating contracts choose not to impose accreditation requirements, the ratings could serve as a point of pressure in Congress and the public at large. Thus, NGOs should spend at least as much energy developing accreditation regimes as they do pursuing transnational litigation under various formal international law instruments. International organizations could also seek to create accreditation regimes. Such accreditation would likely be influential over time, even if states at first formally refuse to implement accreditation requirements into their contracts.

F. Mandating Contractor Self-Evaluation

Contractors could also be required to perform self-evaluations as a way of enhancing accountability. Presented with an internal self-evaluation, an outside monitor, whether governmental or third-party, can often scrutinize the contractor’s performance more quickly and efficiently. Of course, self-evaluation gives the contractor discretion to massage the data and indeed can be subject to outright manipulation and abuse. But nonetheless, it can be a useful starting point for outside monitors, who can at least at the outset make a faster assessment as to whether the contractor has met the contract goals. In addition, self-evaluation can encourage more effective internal policing by the contractor.

Due to these potential benefits, self-evaluation has emerged as a frequent tool in the domestic context. In the world of private prisons, for example, contractors regularly are subjected to self-evaluation requirements. In Texas, prison contractors must “establish performance measures for rehabilitative programs and develop a system to assess
achievement and outcomes.” Likewise in the field of health care, a health maintenance organization must, if it is to receive accreditation, conduct continuous “quality improvement,” in an ongoing internal self-evaluation process. Contracts that require accreditation thus effectively mandate such self-evaluation.

In the foreign affairs context, private foreign aid providers operating under agreement with USAID are regularly required to perform self-evaluation, but foreign aid contracts provided through other agencies and military contracts seem to be devoid of such provisions. Again, taking the publicly available Iraq contracts as an example, none requires the private contractor to file self-evaluation reports, develop internal assessment practices, or otherwise engage in self-evaluation. And while self-evaluation on its own is unlikely to significantly improve contract compliance, such self-evaluation can be useful in combination with some or all of the other contractual provisions discussed in this Article.

G. Enhancing Governmental Termination Provisions and Allowing for Partial Governmental Takeover of Contracts

Contracts could also include terms allowing the relevant government (or international organization) to take over the contract by degrees before ultimately terminating the agreement for failure to observe provisions implementing public law values. Currently, most contracts have implied or explicit provisions allowing only for outright termination for noncompliance. On its face this sort of termination provision seems as if it would provide a strong incentive for contractor compliance. In actual practice, however, outright termination is such an extreme measure that governments are often reluctant to invoke it, and because contractors know that termination is so unlikely, the provisions have almost no disciplining effect.

Thus, it would be better if such termination provisions were supplemented with more graduated penalties, such as provisions permitting the partial governmental takeover of contracts. Because graduated penalties are less extreme than outright termination, they are far more likely actually to be invoked by contract monitors, making them a more effective enforcement mechanism than the harsher (though rarely invoked) termination provisions. Moreover, if partial takeover fails to stem the abuses, outright termination still remains a penalty of last resort. In the domestic context, states are increasingly turning to mechanisms such as graduated penalties, for example, to increase oversight of private nursing homes receiving public funding. Scholars and practitioners have also called for the use of such penalties in the private prison setting.

Turning to foreign affairs, while contracts subject to the FAR do contain termination provisions, they are rarely exercised and are not supplemented by lesser, graduated penalties. As a result, the government has little leverage over contractors. The Iraq contracts provide a notable demonstration of this problem. When CACI employees were implicated in abuses at Abu Ghraib prison, for example, the U.S. government did not terminate its contract. Indeed, although the particular employees implicated in the abuse charges no longer work at CACI, it is unclear whether government actors even so much
as stepped up their supervision of the contracts. To the contrary, CACI actually received a contract extension for interrogation services.

Obviously, governments (and international organizations) should be encouraged to invoke termination provisions when contractors fall short. But even without full termination of the contractors, graduated government (or international organization) takeover could provide an added incentive for contractors to promote public law values.

**H. Allowing for Beneficiary Participation or Broader Public Involvement in Contract Design**

Contracts could also permit beneficiaries or the broader public to help shape contract terms and evaluate performance. In the domestic context, commentators have suggested that such beneficiary participation or involvement by the broader public could greatly enhance the extent to which contractors fulfill public law values. Indeed, as Fred Aman has argued, precisely because privatization contracts are difficult to terminate and sometimes become “immutable,” it is “important that the participation of the public and the public’s representatives be maximized as early in the process as possible.” He thus advocates allowing the broader public to play a role in the design of the contracts themselves.

Some state and local governments have begun to do so. For example, Wisconsin’s contracts with managed care organizations to provide health care to Medicare and Medicaid recipients include provisions for participation by community groups. Other states have gone even further and now require broad public involvement in virtually all privatization decisions. In Montana, for example, any privatization decision must be made subject to a plan available to the public and open to public comment. Other states have similar provisions.

Foreign affairs contracts might benefit from this approach. Indeed, such participation may be particularly important to promote public law values because the ordinary democratic process open to those experiencing the effects of privatization in the domestic context is essentially unavailable for non-citizens outside the United States who are affected by the activities of contractors. To be sure, even in the domestic context, there has long been a worry that privatization removes a crucial democratic check on government. The link between those affected by government action and the government actors is attenuated when that activity is farmed out first from legislatures to agencies, and then from agencies to private contractors. Scholars and policy-makers worry that this form of delegation reduces transparency, which in turn reduces the ability of those affected to vote their preferences when things are not going well. But when governments turn to private contractors to perform foreign affairs functions, the problem is increased exponentially because many of the people affected by the contracts in question do not belong to the U.S. democratic polity or indeed any democratic polity at all. Moreover, U.S. citizens may be less inclined to use the democratic process to voice their views when the effects of contracting are felt mainly overseas.
While it may make less sense to allow involvement of those non-citizens affected by military contractors overseas, due to obvious security concerns, beneficiary involvement or broader public participation in the design and evaluation of foreign aid contracts might be particularly useful. Governments providing long-term development aid through private organizations have to some degree already begun to adopt this approach. In the United States, USAID has allowed local beneficiaries and NGOs to help design development aid agreements, usually on an informal basis, and most frequently when such agreements are negotiated through field offices. Agencies other than USAID, however, are less likely to engage in such consultation. Humanitarian aid and post-conflict reconstruction assistance are also less likely to incorporate such an approach, though recently the UNHCR has begun to explore the possibility of refugee and internally displaced person evaluation of humanitarian aid.

I must leave to a future article a more detailed discussion of how best to maximize opportunities for those affected by a foreign aid project to participate in the design of that project. Certainly, the idea raises a whole host of practical problems. For example, it will be difficult to determine who exactly can speak for an affected population. Is NGO participation sufficient? How does one determine which civil society actors are most representative? What if different sectors of the population disagree as to the efficacy of a proposed project? Even assuming one determines the appropriate voices, what form should the feedback take? Is informal consultation enough? Or should there be a more formal notice and comment period? Or is it necessary to establish an independent tribunal with the power to quash the project altogether? And should such a tribunal be governmental or private? While these questions certainly must be addressed, it seems to me that if we are asking them, we will have already advanced the debate quite a bit. The important point for now is that we must at the very least begin to explore ways of involving in the contracting process itself those affected by foreign affairs agreements. Explicit contractual requirements would go a long way toward facilitating consultation with beneficiary populations, thereby effectuating through contract a broader form of public participation.

I. Strengthening Enforcement Mechanisms

Finally, the contracts could provide for enhanced enforcement mechanisms. They could, for example, give beneficiaries the opportunity for privatized administrative hearings. Additionally, contracts might include third-party beneficiary suit provisions, empowering contract beneficiaries or other interested parties to sue in domestic courts for breach of contract. And whistleblower protections might be enhanced. All of these measures would likely increase compliance with contractual terms.

In the domestic context, governments and policymakers have begun to implement such measures, though private grievance procedures remain more prevalent than broader third-party beneficiary suit provisions and whistleblower protections. Commentators regularly call for an expansion of third-party beneficiary suit provisions (which courts generally refuse to imply unless clearly specified in the contract), but such provisions remain rare. Many private contractors providing aid, however, do offer individual complaint mechanisms for affected beneficiaries. Although these aid providers are not state actors

856
Draft — do not distribute!
and would therefore generally be immune from constitutional review, such contractual provisions do allow for notice and opportunity to be heard, thereby incorporating elements of constitutional due process. These private grievance systems are perhaps most evident in contracts with private prison operators, which typically require such mechanisms. But they appear in other contexts as well, such as health care. For example, the Medicare statute requires that health maintenance organizations receiving federal funding to cover their treatment of Medicare beneficiaries must “provide meaningful procedures for hearing and resolving grievances between the organization . . . and members enrolled.”

Governments might experiment with similar measures in the foreign affairs arena. The World Bank has taken steps in this direction, by enabling aid beneficiaries to bring grievances before special tribunals challenging gross abuses. Third-party beneficiary suit provisions, however, are virtually non-existent, and none of the Iraq contracts contains such a provision. Whistleblower protections should also be enhanced. Government officials currently receive whistleblower protection for reporting abuses in the negotiation or management of contracts, but employees of private companies are not protected under the general Whistleblower Protection Act. In specific statutes, however, Congress has at times extended whistleblower protection to private employees. For example, seven of the major federal environmental statutes contain such whistleblower clauses. Thus, whistleblower protection could also be extended to private sector employees working for a government contractor, who provide information concerning the unlawful performance of a contract. Such a provision, combined with the availability of third-party beneficiary suits, or possibly even qui tam actions, would go a long way towards making sure that any contract-based efforts to provide accountability will have back-end enforcement to encourage compliance.

As discussed previously, enforcing international law norms through contract, rather than directly in an international forum, obviates any need to argue that the contractor should be deemed a state actor. In addition, because international law enforcement mechanisms are relatively weak compared to their domestic counterparts, a contractual approach is far more likely to lead to meaningful judicial review. In addition, requiring domestic judges to enforce international public law values embodied in contracts may have important norm internalization effects because such judges would essentially be enforcing international law norms. This increased familiarity with international law principles might lead to less resistance to those norms as a general matter, thereby effectively expanding the reach of international law.

On the other hand, one might argue that localizing the enforcement of international law norms might either cause international enforcement mechanisms to atrophy from disuse or lead to heterogeneity in different countries’ understandings of the principles, which could undermine the notion of a common international law. Neither objection, however, should create serious hesitation about pursuing contractual accountability. First, as previously discussed, international law enforcement mechanisms are already weak, and to the extent that they have been effective, at least in the human rights context, it has been by selectively limiting the scope of enforcement to the very most egregious human rights violators. Thus, it is not at all clear that providing a possibly effective domestic avenue
for pursuing claims against private actors (who would have been unlikely to face proseuction internationally in any event) will in any meaningful way undermine international law institutions. Second, to the extent that domestic judicial systems, government officials, and broader populations internalize international law norms, it strikes me that the benefits of such norm internalization far outweigh any possible concern about maintaining the “purity” of the international norm. Local variation is to be expected, of course, but such [heterogeneity] in the domestic incorporation of international norms strikes me as a strength, not a weakness. Finally, the key point is that without focusing on contracts, there may be no realistic way to impose norms of accountability on privatized foreign affairs activity at all. Accordingly, those seeking to expand the applicability of international law norms should at the very least seriously consider using contractual enforcement mechanisms or risk the possibility that such norms will simply be ignored in an increasingly privatized world.

IV. Conclusion

Resisting privatization in the foreign affairs context is probably no longer an option. Indeed, if anything the scope and pace of privatization in the international arena is increasing. Moreover, it will not be sufficient merely to tweak existing international law treaties or doctrines (or even invent new ones) in order to bring private contractors within the ambit of formal international law. After all, even if international or domestic courts could be convinced that private contractors should be held liable for violation of international law norms (which is far from certain), international and transnational public law litigation will never be able to hold accountable more than a handful of people. Accordingly, those who seek to preserve or expand the values embodied in public international law will also need to look elsewhere to find mechanisms for ensuring accountability in a privatized world.

In this Article, I have suggested one such mechanism: the government contract that creates the privatized relationship in the first place. Drawing on the far more extensive domestic administrative law literature on the subject, I have identified a variety of provisions that could be incorporated into such contracts. These provisions seek to encourage compliance with (and enforcement of) human rights and humanitarian law, ensure transparency and democratic accountability, and promote norms against corruption, waste, and fraud. Taken together, they provide a menu of options for regulators, activists, policymakers, and scholars who are concerned at the potential for abuse in our current contracting processes.

Of course, governments may be hesitant to insist on some of these contractual provisions. For example, officials may fear that such requirements could unduly increase the costs of privatization both to the contractor and to the government entity overseeing the contract. Or, more cynically, resistance might stem from the fact that governments actually benefit from a more opaque process with less public oversight. In any event, one seeming difficulty with relying on contractual provisions is that the increased oversight will be included in contracts only as a “matter of legislative or executive grace,” and therefore can be rescinded or limited at any time.
Yet, such objections do not render a contractual approach unrealistic. To begin with, concerns about the cost of additional contractual requirements may well be over-stated. As the Custer Battles fiasco makes clear, in many cases better oversight could actually save the government far more money than it costs. And as to concerns that added contractual provisions will cause contractors to walk away or prohibitively raise their rates, the short answer is that far more empirical work must be done to assess whether such dire predictions are accurate. After all, it seems quite unlikely that contractors bidding for these extraordinarily lucrative contracts with governments such as the United States will pull out of the process just because of some added contract requirements. To the contrary, the government should, by all rights, have tremendous leverage in the contracting process because there are unlikely to be competing customers similarly able to offer billions of dollars in contract awards. Indeed, while government contractors in the past have often raised concerns about increased compliance costs to object to enhanced contractual oversight, at least one commentator has challenged such claims, noting the absence of compelling evidence that increased oversight through, for example, qui tam suits has resulted in a significant number of firms refusing to do business with the government.

In addition, while some governmental officials surely would prefer a more opaque process, governments are not monolithic entities, and proposals such as the ones outlined in this Article may be taken up and championed by members of the bureaucracy, even without the imprimatur of higher level executive branch officials or the legislature. Moreover, it is incorrect to think that more robust contractual monitoring can only come about through official executive branch or legislative action. First of all, some of the proposals for monitoring of contracts and accreditation or rating of contractors could be undertaken by NGOs or other groups without any official action whatsoever. While such evaluations might not initially have the power of the state behind them, the example of NCQA indicates that, over time, governments can be convinced to adopt a previously unofficial rating system as its own. Second, even if governments never adopted the standards, simply the process of evaluating and accrediting contractors would provide a rich source of public information about privatization that could be used to bring popular political (or economic) pressure to bear on noncompliant contractors. Such public reporting might also allow citizen watchdog groups (or even competing contractors) to monitor the effectiveness of particular contracts, publicize deficiencies, and lobby government officials for change. Third, advocacy at the international level could result in treaties or other international regimes that actually require governments to include oversight provisions in certain categories of contracts, thus creating increasing pressure for change. In any event, as the domestic examples demonstrate, governments and agencies can, at least at times, be mobilized to require meaningful contractual oversight.

In the end, whatever the drawbacks of a contractual approach, they are certainly no greater than the weaknesses of the existing formal transnational/international court system. Indeed, the use of contractual provisions has the benefit of opening up the possibility of legal enforcement regardless of whether or not there is state action and to provide the foundation for legal action in domestic, as well as international, fora. Such contractual mechanisms might also pave the way for statutes and treaties. Thus, international law scholars, activists, and advocates should spend at least as much time
studying and lobbying for contract-based compliance regimes as they do seeking further openings for international or transnational litigation.

Perhaps most importantly, we must remember that the proper management of privatization will almost certainly require a variety of approaches, and we need not choose one to the exclusion of others. My aim here is simply to focus attention on privatization in the international realm as a crucial field of study, to call for dialogue among international and domestic scholars, advocates, and policy-makers concerning appropriate responses, and to suggest that more attention be paid to the possibility of using contractual provisions to provide accountability. None of these aims requires that contract become the only response to privatization. To the contrary, in the coming years we will need to think broadly and creatively about how best to respond to the threats posed by the outsourcing of governmental functions to non-governmental entities. Only through such efforts will we be able to find ways to protect crucial public law values in the era of privatization that is already upon us.


I. INTRODUCTION

In late 2002, while grabbing headlines for boldly promising to slash the federal civilian workforce in half, the Bush Administration was at the same time discreetly hiring private contractors to relieve Special Forces troops of their duty to protect President Hamid Karzai in Afghanistan. In the more celebrated declaration regarding workforce reductions—perhaps the culmination of a decade-long, bipartisan initiative to reinvent and streamline government—the President attempted to allay concerns by stressing that the proposed job cuts would not intrude on any functions that are “inherently governmental;” these cuts would instead be focused more narrowly on reaping economic benefits by privatizing commercial responsibilities such as catering, gardening, and clerical work. Unfortunately, in replacing Special Forces troops with private military contractors, the Administration offered no comparable words of comfort.

Since then, although the government has subsequently scaled back its ambitious domestic downsizing and privatizing initiatives, it nevertheless has expanded and intensified its military privatization agenda. This has especially been the case in Iraq, where today over 20,000 contractors are securing key American installations, participating in armed raids against insurgents, and—most infamously—serving as interrogators in the occupation’s most notorious prisons.

Who would have thought that when the modern wave of government privatization began decades ago with cities experimenting with the contracting out of their sanitation responsibilities, it would swell to encompass the privatization of prisons and welfare
services, let alone the privatization of foreign policy and national defense? Even staunch libertarians, proponents of the Nozickean night-watchman’s state, have long-conceded that when stripped to its core, a nation still must maintain its public commitments to national defense. Indeed, just a few years ago, leading privatization scholars dismissed as implausible the idea that we privatize national security functions.

These individuals, like many others, would thus not have expected Washington—over the past decade under Democrats and Republicans alike—to employ private agents to do its military bidding in the Latin American drug wars, the Balkans, the Middle East, Rwanda, Afghanistan, and now, in Iraq. In short, since the first Persian Gulf War, private soldiers working for “military firms” under contract with the U.S. government have seen active duty in most conflicts involving the United States (and also some in which the United States has had no official military involvement). In another era, we would call these agents “mercenaries” and label their sponsor governments immoral and illegitimate; could it be that, today, these actors are just another set of government contractors, and the United States is just outsourcing one more governmental function?

Observers who react with dismay over the outsourcing of military functions might see it as the modern, or perhaps post-modern, embodiment of President Eisenhower’s famous warning in 1961, when the former Supreme Allied Commander portended the rise of the military-industrial complex. But while Eisenhower’s prescient words continue to resonate today—as we witness the awarding of hundreds of contracts to private firms, often to those quite friendly with high-ranking government officials, to rebuild the infrastructure and restore the institutions of Iraq and Afghanistan as well as scores of additional contracts for defense hardware—even he could not have foreseen the government’s current policy of delegating highly sensitive responsibilities to private soldiers in and near zones of conflict.

Indeed the delegation of combat responsibilities presents a qualitatively different and more dangerous privatization agenda than that which troubled Eisenhower. His concerns would be reflected today in the recent allegations of “sweetheart” deals between the federal government and the likes of, say, Halliburton for energy services in Iraq or Boeing for Tanker aircraft. But the harms that flow from those types of contracts, however troubling and possibly even scandalous, fit comfortably within the conventional privatization framework of outsourcing functions that are not inherently governmental, but rather are commercial in nature. They are problems of accountability, and result mainly from poor oversight, improper contract management, and insufficient fidelity to (or simply inadequate) conflict-of-interest laws. And although these contracts and the harms that may accompany them are worrisome from an array of policy perspectives, conceptually speaking they are unremarkable: Driven by the same market-efficiency impulses that motivate the outsourcing of sanitation, catering, and even prison management responsibilities, the contracts to rebuild roads and schools in failed states and to manufacture new weapons do not compel us to rethink our basic understandings of American privatization.

Military privatization of combat duties, on the other hand, decidedly does. It has the potential to introduce a range of novel constitutional, democratic, and strategic harms that
have few, if any, analogues in the context of domestic, commercial outsourcing. Military privatization can be, and perhaps already has been, used by government policymakers under Presidents Bill Clinton and George W. Bush to operate in the shadows of public attention, domestic and international laws, and even to circumvent congressional oversight. For a variety of political and legal reasons, the Executive may at times be constrained in deploying U.S. soldiers. The public’s aversion to a military draft, the international community’s disdain for American unilateralism, and Congress’s reluctance to endorse an administration’s hawkish foreign goals may each serve to inhibit, if not totally restrict, the president’s ability to use U.S. troops in a given zone of conflict. In such scenarios, resorting to private contractors, dispatched to serve American interests without carrying the apparent symbolic or legal imprimatur of the United States, may be quite tempting.

In those instances, it would not necessarily be the cheaper price tag or specialized expertise that makes private contractors desirable. Rather, it might be the status of the actors (as private, non-governmental agents) vis-à-vis public opinion, congressional scrutiny, and international law that entices policymakers to turn to contracting. Indeed, “tactical privatization,” as I call it, is motivated at least in part by a desire to alter substantive policy: Private agents would be used to achieve public policy ends that would not otherwise be attainable, were the government confined to relying exclusively on members of the U.S. Armed Forces. Tactical privatization thus stands in contradistinction to what is widely understood to be the conventional privatization agenda, driven by economic goals, that strives for verisimilitude in replicating government responsibilities (only more efficiently). To elude public debate, circumvent Congress’s coordinate role in conducting military affairs, and evade Security Council dictates may help an administration achieve short-term, realpolitik ends; but in the process, the structural damage to the vibrancy and authenticity of public deliberation, to the integrity of America’s constitutional architecture of separation of powers, and to the legitimacy of collective security may prove irreparable.

What is perhaps worse, the structural harms introduced by decisions to privatize may not substantially lessen even if, or when, combat privatization is undertaken relatively transparently and mainly for more traditional, commercial reasons. Since much of Congress’s chief warmaking powers flow from its legal authority over the Armed Forces (especially to authorize armed intervention), even assuming the aims of privatization are purely economic and unconnected to any tactical motives to subvert Congress, constitutional harms do not disappear. In those situations and however inadvertently, privatization would still circumscribe Congress’s role in military affairs, thus prompting separation-of-powers concerns not altogether dissimilar to those that would exist were the circumvention intentional. Additionally, and also irrespective of the Executive’s motives for privatizing, the introduction onto the battlefield of for-profit contractors, motivated to fight primarily by money and regulated loosely by contract, rather than by the Uniform Code of Military Justice, breeds an array of strategic and psychic harms for the military commanders, for uniformed soldiers in the field, and for Americans at home. Accordingly, privatization of military functions poses a slew of problems too complicated and varied to resolve merely by enhancing accountability, strengthening contract laws, and tightening contract management.
It is, therefore, the present aim of this Article to identify in yet unexplored ways the profound and pervasive dangers that this new modality of privatization introduces. To date, commentators writing about military privatization have primarily focused on the tangible misdeeds that privateers have perpetrated in zones of conflict and on the reform measures necessary to improve battlefield accountability. But what these scholars have overlooked are the deeper, structural problems. Accordingly, this Article seeks to look beyond economic efficiency and accountability concerns—the principal foci of privatization scholarship—to explore how covert and, at times, even transparent delegations of sensitive military responsibilities threaten to (1) violate the constitutional imperatives of limited and democratic government, (2) undermine the institutional excellence of (and patriotic support for) the U.S. Armed Forces, and (3) jeopardize the already shaky diplomatic and moral standing of the United States in the eyes of the rest of the world.

This Article proceeds in six parts. I begin in Part II first by tracing the modern evolution of military privatization and next by discussing six contemporary case studies. Then, I attempt to locate some of the normative impulses motivating this new wave of privatization and to situate them within the broader pattern of American privatization policy; this last section serves to frame the principal conceptual differences between combat-related and more conventional forms of privatization, which will be important in understanding the unique structural harms introduced by decisions to outsource military responsibilities.

In Part III, I commence with the inquiry’s critical analysis: understanding these structural harms. In this Part, I describe how the Executive can use military contractors to direct national security policy with greater impunity and less oversight than it could if it only had U.S. troops at its disposal. To the extent that Congress’s war-making authority is tied primarily to its regulatory and war-authorizing powers over the American military qua U.S. Armed Forces, a president interested in exercising more unilateral control might hire private contractors in lieu of U.S. soldiers and hence avoid having to collaborate as closely with the legislative branch. In circumventing congressional authority, the Executive violates the two principal constitutional imperatives: limited government—by bypassing Congress and preventing it from checking the ambitions of the president—and democratic government—by acting covertly (i.e., without congressional or, by extension, the People’s input) and thus failing to make inclusive policy decisions legitimated by popular consent. While a paradigm case of tactical privatization would involve executive intent to evade congressional monitoring and to avoid having to request authorization for engaging troops in hostilities, harms along these lines would nevertheless ensue even if the president had no such insidious objective—and was instead focused mainly on maximizing economic efficiency. Simply and even inadvertently operating outside of the carefully arranged framework of coordinate military policymaking over the U.S. Armed Forces still has the effect of limiting Congress’s formal and informal involvement in decisionmaking.

Then, in Part IV, I characterize how the introduction of private troops, either integrated into a larger contingent of U.S. military personnel or instructed to operate independently, creates considerable institutional harms, strategic liabilities, and morale problems. First,
because privateers are not bound by the dictates of the Uniform Code of Military Justice, but rather often only by the terms of their contract, there is a much greater likelihood that they might abandon or distort a mission, ultimately prioritizing some economic goal or their own personal security over the task at hand. Importantly, I argue that this harm goes beyond mere battlefield accountability concerns since it is not so much the potential for privateers to botch a mission that represents the foremost problem; rather, because contractors cannot be regulated as stringently as U.S. troops, also at issue here is the legal dilution of military justice and discipline, on and off the battlefield. The contractors’ presence, their uncertain legal status, and their relative impunity from courts-martial could destabilize the delicately balanced constitutional arrangements associated with civil-military relations and democratic warmaking. And, second, I examine how the presence of contractors (to the extent they are publicly perceived as profit-seekers rather than as patriots) on the same hostile terrain as regular soldiers may ultimately threaten the privileged and honored status the military has historically enjoyed among the American public.

In Part V, . . . I discuss the international/diplomatic harms privatization engenders. I describe how military privatization can exacerbate foreign critics’ worst fears and suspicions about the United States: No longer will the United States retain the moral high ground by risking its own young men and women of a volunteer army in the name of freedom. Instead, a critic assumes, outsourcing gives Washington freer rein by allowing the government to indemnify itself against casualties and other “sticky” political situations and therefore permits it broader license to purchase strategic outcomes. Moreover, privatization, to the extent that it allows the United States to bypass international agreements and Security Council authorization, undermines the legitimacy and vitality of collective security. Although these harms are felt primarily by the outside, non-American world, they nevertheless have adverse consequences for American foreign policy, for American integrity, and for the interests of containing and regulating the proliferation of even more odious strains of military profiteering that exist in other parts of the world. Therefore, I argue, these international implications should weigh heavily on any structural assessment of the virtues and vices of using private soldiers. Note that whereas the harms explored in Part III chiefly occur when Congress’s role is subordinated, the problems analyzed in Parts IV and V do not necessarily depend on circumventing congressional participation in military privatization.

Part VI concludes by first roughly sketching out a set of reform measures that might help to reduce the legal and symbolic status differentials between contractors and soldiers that underlie many of the manifest structural harms described above. Having proffered some reform proposals, I then consider which status disparities may prove the most difficult to eliminate. Finally, I discuss whether these reforms, if successful, might actually reduce, if not altogether destroy, military privatization’s raison d’etre.
II. THE MODERN AMERICAN EXPERIENCE WITH MILITARY PRIVATIZATION

C. Conceptualizing Tactical Privatization

... Without having looked at [recent examples of military privatization], one might readily assume that economic efficiency, the *sine qua non* of privatization, explains the evolution and expansion of military outsourcing. But these examples, which paint a vivid though still inchoate and fragmentary picture of military privatization, actually suggest that there might be alternative (or at least additional) reasons why policymakers employ contractors. ... We begin to realize that not only must we grapple with the implications of the dynamic transformations from outsourcing strictly commercial functions to ones involving the exercise of considerable discretion of the sort normally considered "inherently governmental;" we must also come to terms with the possibility that conventional, economic justifications do not explain the full breadth of normative reasons for soliciting private soldiers.

Traditionally, the lens of privatization scholarship has focused on economic efficiency, how competitive market forces and profit incentives can inject cost-savings and quality-enhancing measures into the provision of government services and functions. Scholars have also examined ways in which contracting out may generate additional cost-savings benefits. For example, contractors are not subject to the costly and time-consuming notice-and-comment requirements of the Administrative Procedure Act or to the disclosure mandates of the Freedom of Information Act. Nor are they necessarily deemed “state actors” for purposes of *Bivens* or 42 U.S.C. § 1983 liability. Finally, employees of contracting firms are less likely to have union protection, and thus they can be made more responsive to market incentives (and more easily fired) than can civil servants. Accordingly, the lower costs associated with contracting out are thus a function not only of competition and innovative business planning, but also of public-private status differentials. Even though they provide cost-savings too, these incidents of privatization, which permit contractors to bypass channels of accountability and to use more “casualized” labor, are, especially since the government is outsourcing increasingly sensitive functions, a growing source of concern.

In the military context, non-economic status differentials can emerge as all-important in (rather than incident to) decisions to privatize. Private actors *qua* private actors may be sought—not because they are situated in a more efficient market or even because they command lower market wages, but because legally, politically, and symbolically they are not soldiers. Military privatization can allow the government to achieve national security and even humanitarian ends that would be more difficult, if not impossible, to accomplish using American soldiers. Perhaps, at various times, a desire, however latent, to avoid instituting a draft, to lessen public awareness, to dilute casualty counts, to bypass congressional troop limitations, and/or to evade international arms embargoes, entice policymakers to outsource because private actors are not regulated, controlled, or even mourned to the same extent that public soldiers are. But, if a decision to outsource does reflect “tactical” aims to circumvent political and legal obstacles associated with the conventional deployment of regular, U.S. troops, an entire set of problems for constitutional principles and democratic virtues—*independent of any actual, tangible*
misdeeds that privateers may perpetrate in a zone of conflict—must be anticipated. It is these structural problems, deeper than just accountability concerns, which command my attention. Indeed, these structural problems are so great in the context of military privatization that even absent any express intent by the Executive to leverage or exploit status differentials between contractors and soldiers, many of the chief constitutional and democratic harms would still arise.

Economic privatization is, ostensibly speaking, ideologically agnostic. Its advocates may have particular agendas, but efficiency-driven privatization per se mainly creates an alternative process for carrying out government contracts that strive to replicate government provision—only at a fraction of the cost (and perhaps with less government red-tape). On the other hand, “tactical” privatization, which may seek to exploit status differentials, is predicated on substantive rather than administrative or bureaucratic reform. Privatization, in this latter case, could be used to achieve objectives materially different than those that could be—for a number of reasons—achieved within the public sector. For example, a conflict may prompt an outsourced response if it would otherwise be difficult for the president to secure congressional and/or international support to deploy members of the U.S. Armed Forces. In such scenarios, it is not the cheaper price tag, but rather the status of the private actors (as distinct from U.S. military personnel) vis-à-vis congressional oversight, public attention, and international law that may motivate policy planners to hire contractors.

In this Section then, I focus on the structural challenges posed by forms of military privatization that leverage status differentials, purposefully or even inadvertently. This is not to say that there are not instances where “tactical” aims may influence outsourcing patterns in domestic policy contexts. Nor is it the case that the lessons and insights we can glean from the already rich, nuanced, and comprehensive scholarship on economic privatization would not be immeasurably helpful here. I nevertheless leave much of that conventional, scholarly analysis to the side for now in order to explore some of the deeper concerns that are triggered when privatization can be undertaken for purposes of limiting political and legal oversight. Thus, for instance, I do not consider the potential economic gains and efficiencies associated with military privatization, though such an exploration would, no doubt, prove quite interesting.

Instead, I briefly discuss how the “optics” of privatization as well as how the legal and political differences between using private troops and American soldiers could create opportunities for national security policymaking that would not be possible were the Executive limited to deploying only members of the U.S. Armed Forces. This short discussion, in turn, helps lay a framework for examining in Parts III, IV, and V, how status differentials not only threaten effective service provisions, but also may disrupt the democratic and constitutional workings of the federal government.

1. Using Private Troops To Minimize Political and Legal Contests

As will be explored at length . . . , privatization expands the horizon of executive policymaking discretion in the context of military affairs. Using privateers, whose legal status differentiates them from regular, U.S. soldiers, could help enable the president to
bypass congressional oversight and even international collective security arrangements. Indeed, outsourcing may be undertaken to exploit this legal gap between what is the official state policy (say, non-intervention, limited involvement, or limited troop deployment) and what military goals can actually be accomplished through private channels. If contractors operate within these interstices, the president can presumably satisfy national security aims without expending the time and political capital to secure formal approval at home or internationally.

First, pursuant to the U.S. Constitution, customary practice, and statutory framework laws such as the War Powers Resolution, the president shares many warmaking powers with Congress. While retaining exclusive jurisdiction over command decisionmaking, the president must nevertheless seek, inter alia, authorization and funding from Congress to deploy U.S. troops into zones of hostility. But, many of Congress’s powers over military affairs are keyed to its Article I authority over the Armed Forces per se. Congress can, for instance, regulate the use and number of servicemen and women abroad, curtail funding for operations, and withhold support for a military engagement. Hence, as it stands, the president must often seek congressional approval in some form or another.

If the Executive were, however, to deploy private troops in lieu of U.S. soldiers, it might be able to evade much of Congress’s oversight jurisdiction—at least temporarily. Without having to seek authorization and funds from the national legislature, the president can more easily engage in unilateral policymaking and dispatch private contractors who are not part of the regular U.S. military. In so doing, objectives can perhaps be achieved more swiftly and with less political wrangling and opposition. This privatization agenda is discussed further in Part III.

Second, an additional—and this time constitutionally exogenous—check on presidential discretion comes by way of the United Nations Security Council. In the post-Cold War era, the Security Council has reemerged as a, if not the, legitimate source for the authorization of military intervention in the name of collective security. Without the endorsement of the Security Council, any one nation’s decision to intervene in the affairs of another sovereign state is subject to criticism and charges of illegality and illegitimacy. But although the Security Council attempts to regulate the behavior of nation-states and their national militaries, it (like international law more generally) has comparatively less influence over the activities of private agents.

If a country were to utilize the services of private contractors, it could bypass a Security Council vote—or possibly evade an already passed resolution prohibiting intervention by member states. Thus, the use of private troops in lieu of the U.S. military may free the Executive from having to depend on the support of the Security Council in order to initiate a foreign deployment. This privatization agenda is explored at greater length in Part V.

2. The Optics of Military Privatization

Beyond leveraging the legal status differentials between U.S. soldiers and private actors to evade oversight by Congress (and maybe even the U.N.), the Executive might further,
or alternatively, resort to privateers precisely because they may have a different social or symbolic status in the American consciousness. Privateers do not, so it appears, occupy the same special place in the hearts and minds of the American public as do its citizen-soldiers. By contrast, it is that special regard for soldiers, well-understood by military and political leaders alike, that often constrains the government from readily sending public troops into harm’s way.

Conversely, it is doubtful also that privateers go overseas with the symbolic “baggage” that U.S. soldiers tend to carry—as exemplars (at least in the minds of many) of hegemony and coercion.

Hence in either or both scenarios, the use of private agents may prove more palatable (or at least more discreet) than sending in the Marines. Dispatching private contractors may not trouble and worry the American people as profoundly as if their boys and girls in uniform were sent into battle. And, likewise, dispatching private troops—who do not even wear the uniforms of the U.S. military and are not as likely to hoist an American flag in celebration on foreign soil—may accomplish goals more readily and with less resistance than if U.S. soldiers were actually deployed.

Below I posit that differences between soldiers and contractors, based on (a) normative value judgments and traditional affinities for U.S. servicemen and women, and (b) sensitivities of foreign hosts, may lead policymakers to prefer private contractors in certain situations. The harms associated with exploiting these sets of status differentials will be addressed more fully in Part IV.

a. Public Opposition Grounded in an Expectation of Zero-Casualties: A Focus on Soldiers’ Deaths

Americans’ general distaste for war is a significant factor circumscribing the government’s ability to deploy and use force abroad. But that aversion is not necessarily grounded in pacifist or even isolationist sentiments; another significant factor is a low tolerance for casualties: the squeamishness associated with watching soldiers arrive home in body bags and with tallying the rising casualty counts in the morning newspapers. Indeed, though the United States has not necessarily been shy about military interventions in principle, it has often been hyper-vigilant about minimizing soldiers’ casualties in any way possible. Billions of dollars expended for stealth fighters, cruise missiles, unmanned drones, and smart bombs aim to ensure that harm to American soldiers is kept to an absolute minimum. In fact, key military decisions are at times made with the public’s concerns in mind even at the expense of sound national security policymaking. For instance, in his efforts to galvanize domestic support for intervening in Kosovo, President Clinton publicly and repeatedly promised not to engage in a ground war. His pledge not to put troops in harm’s way may have secured the public support at home necessary to liberate Kosovo, but it also reduced the strategic discretion the Pentagon would have otherwise possessed were no such promise made.

An attitude of risk-aversion and faith in what is the now-popularized “zero-casualty,” force-protection military paradigm constrains the effective exercise of military power—
but not as much as if the overriding concern among Americans were purely pacifist in nature. This distinction between a zero-casualty and pacifist mentality may be less meaningful in the context of sending American troops into a conflict zone: either way, the public would be reticent to support a combat-related engagement. But, in the context of employing private troops who may not have the preternatural connection to the American people that U.S. soldiers enjoy, this distinction might make all the difference in how a president conducts foreign policy.

Enter the contractor. Tim Spicer, founder of Sandline, a prominent British military firm, believes military contractors can “fill the gap” left in the wake of the debacle in Somalia more than a decade ago. Recall that America’s low tolerance for casualties, perhaps a by-product of Vietnam, was tragically tested in Somalia, where the sight of American soldiers dragged through the streets of Mogadishu was televised stateside for all to see. Indeed, “the live footage on CNN of United States troops being killed in Somalia has had staggering effects on the willingness of governments to commit to foreign conflicts.”

Private firms can undertake dangerous missions on behalf of the U.S. government without the attention, media coverage, or official sponsorship; if things go wrong, the line of blame to the government is more attenuated and the casualties would not be patriotic American soldiers serving under (and being carried home under) the American flag, but rather defense contractors whose deaths are not officially reported. As former U.S. Ambassador to Colombia Myles Frechetter noted: “If the narcotraffickers shot American soldiers down, you could see the headlines: ‘U.S. Troops Killed in Colombia.’” But when three DynCorp employees were shot down during an anti-drug mission in Peru, their deaths “merited exactly 113 words in the New York Times.” And, as Doug Brooks, a private military industry spokesperson explains, if an American soldier is killed overseas, it is front-page news. If he is not a soldier, and instead is a private contractor who “is shot wearing blue jeans, it’s page fifty-three of their hometown newspaper.” Journalist Kevin Myers has come to a similar conclusion: If a private military contractor is “killed in action, the tabloid sob-industry cannot then move into tearful action, wondering about our brave boys perishing on a foreign field.” In the hearts and minds of the people, private actors “are excluded from such hand-wringing.” Indeed, although ABC’s Nightline recently devoted an entire episode to a solemn reading of the names of the slain American servicemen and women in Iraq, it is highly doubtful that it or a similar show would allot comparable time to fallen contractors.

Thus in conflict zones, or areas of potential conflict, such as Colombia, Afghanistan, Iraq, and Rwanda, the use of private agents rather than American soldiers does not lower the likelihood of death. But their acting in lieu of soldiers does perhaps lower the likelihood of the unacceptable imagery of American soldiers coming off cargo planes in bodybags draped with the flag. It is possible that, at least in terms of small-scale operations (such as in Rwanda or Sudan), this gives the president greater discretion to place troops on the ground for humanitarian peacekeeping or even hostage-rescue assignments that the public would deem too remote an interest to justify jeopardizing American soldiers. And, in high-profile interventions, such as in Iraq or Afghanistan, the use of contractors can lower the number of soldiers who have to be called into or kept in service, dilute the tally of official casualties, and lessen the need to cultivate a broader international coalition.
The U.S. government may, in turn, exploit this gap in how contractors are valued vis-à-vis soldiers and place privateers in harm’s way at a lower political cost.

Perhaps these observations overstate the difference, especially in light of the Bush administration’s ongoing War on Terror and the war in Iraq. After September 11, the force-protection theory of war-making may seem more of a naïve luxury than a sustainable national defense strategy. Casualties to American troops struck down in the caves of Afghanistan or the streets of Iraq may be considered acceptable in ways they might not have been in Kosovo or Colombia. And, in Iraq, as contractors become more commonplace on the battlefield and more closely associated with the American commitment there, the symbolic differences between them and soldiers may lose some currency. Accordingly, to the extent the differences lose meaning, however, so does the policymakers’ perceived flexibility to employ privateers as less politically costly stand-ins (and hence contractors may become less useful).

Nevertheless, the public’s sense of the differences may endure—and may even become more acute in instances where national security interests are not implicated. In other words, the loss of military lives in a humanitarian intervention—conducted contemporaneously as “real” wars are being fought on the frontlines of American security interests—may become even less acceptable. But, as often is the case with trying to glean meaning from dynamic trends, this discussion is speculative, of course, and any statements proffered here would benefit from further empirical analysis and/or a longer period of time to gauge cultural changes brought about in the post-September 11 climate.

b. Lowering the American Profile Abroad

Moreover, at times U.S. expertise and strength may be warranted—and even solicited by foreign leaders—but the symbolism of inviting American troops may prove too problematic for the host country, and the decision to dispatch them may do more harm than good. One need only consider the level of hostility shown toward U.S. GIs in countries with complicated histories of an American military presence, such as Japan, Saudi Arabia, and the Philippines, to appreciate that in some circumstances private contractors not wearing uniforms and not waving American flags may be much more effective agents of foreign policy than would soldiers, whose presence often invites anti-American sentiments.

Contractors, even if they are all Americans, may not exhibit any telltale signs of nationality. Hence, they may be especially valuable in places where the willingness of foreign leaders to help the United States fight the War on Terror exists but is offset by strong domestic opposition to U.S. forces on the ground. After all, one of Osama bin Laden’s principal reasons for threatening Saudi Arabia remains the Kingdom’s willingness to host American military bases in the “Holy Land.” And, on the flip side, the deaths of American contractors overseas (as opposed to U.S. soldiers) may be less likely to lead to a public outcry at home, which then might require the United States to respond with even greater force in defending its interests.
III. Threatening the National Security Constitution

While the immediate benefits of cost-savings, economic efficiency, and greater political maneuverability provide strong incentives for policymakers to consider employing private contractors, a full accounting of the concomitant harms is also in order. In the parts that follow, I focus on structural harms and catalogue the depth and breadth of the potential dangers brought about when core governmental responsibility over military engagement is delegated to privateers. Indeed, whether explicitly seeking to evade political and legal constraints—or even inadvertently doing so in the course of trying to save money—the enhanced discretion associated with military privatization may: (1) subvert the constitutional imperatives of limited and democratic government, (2) diminish the effectiveness of the U.S. Armed Forces, and (3) undermine the already weak diplomatic and moral standing of the United States abroad.

In this Part, I focus on how private contractors may enable the Executive to conduct military policy with relatively few constraints. To the extent that Congress’s authority over warmaking is principally tied to its Article I powers over the U.S. Armed Forces, a president seeking more unilateral control might deploy private troops instead of U.S. soldiers. By bypassing congressional authority, the president violates the two chief constitutional imperatives: limited government—by circumventing Congress and limiting its ability to rein in the power of the president—and democratic government—by acting covertly without the national legislature’s and, by extension, the People’s consent. Problematically, even if the Executive had no such insidious aim—and was instead seeking primarily to maximize efficiency gains—simply and even inadvertently operating outside of the constitutional framework of shared military policymaking has the effect of limiting Congress’s and, again, the People’s formal and informal involvement in national security affairs, a limitation that, of course, is harmful to the proper functioning of government. For the most part, over time, Congress should be able to recover and reassert much of its authority by actively legislating to impede or, perhaps just counterbalance, the president’s unilateral activity. Therefore, presidential discretion by way of outsourcing may not create an insurmountable constitutional crisis, but can, at the very least, create a critical imbalance that has yet to be satisfactorily anticipated.

And, in the subsequent two Parts, I discuss, first in Part IV, how military privatization damages the institutional integrity and effectiveness of the U.S. Armed Forces and, also, how it may threaten the normative standing of the American soldier as an embodiment of the patriot-citizen; and then in Part V, I characterize how military privatization, by undermining the legitimacy and vitality of collective security agreements, provides additional fodder for those already suspicious of American foreign policy.

Some of these harms identified in this Part as well as Parts IV and V, lend themselves to amelioration through more procedural transparency, through legislation mandating greater coordination with Congress, and through more candor with the American people. Other harms, however, are more intractable and, for constitutional and cultural reasons, not as easily remedied. A discussion of an agenda for reform—and the limitations of reform—will be reserved for this Article’s conclusion.
A. Military Privatization’s Threat to Limited and Democratic Governance

Although we might think of the call to Philadelphia in the Summer of 1787 as a concerted effort to redistribute power away from the national legislature and toward a strong Executive, the Founders nevertheless retained for Congress a sizable bulk of the Republic’s warmaking powers. Scholars have suggested that the motivation for the Convention lied principally in addressing the Articles of Confederation’s defects in domestic governance (as well as in its misallocation of powers between the states and the Union), rather than any shortcomings in the nascent country’s perceived abilities to take up arms in defense of its sovereignty. Indeed, perhaps centuries of Old World tyranny and scores of bloody wars instigated by petulant European kings sensitized the Founders to the dangers of entrusting the sword and the decision to wield that sword in the same set of hands. Vesting warmaking decisions—to authorize war, fund war, and supply and regulate the personnel involved in war—in Congress advanced, as I have intimated above, the two chief aims of the American experiment in constitutional democracy. The United States would be a limited government: its Commander-in-Chief would be constrained by sets of laws, deliberative processes, and by other, equally ambitious leaders in coordinate branches. And the United States would also be a great democracy: its decisions would reflect the will of the citizenry. Hence, Congress[,] as the most direct representatives of the People, would necessarily be involved in military policy, simultaneously promoting the virtues of limited government by checking the perceived natural inclinations of the strong Executive and upholding the ideals of democracy by remaining the true servants of the People. Moreover, decisions by the president to wage war could not be undertaken without first benefiting from the deliberative insights of a legislative assembly and without concomitantly securing the tacit blessings and consent of the citizenry. Military privatization threatens this framework of coordinate decisionmaking. The potential to outsource combat roles necessarily carries with it opportunities for the Executive to wield powers unimaginable were it limited to the use of regular, U.S. troops. By shifting responsibilities away from America’s armed forces and delegating them to private contractors, the president can circumvent constitutional obligations to share warmaking authority with Congress. Privatization, therefore, may destabilize the delicate balance of powersharing built into what Dean Harold Koh calls the National Security Constitution, by weakening a critical check on presidential power—a failure of constitutional governance—and also by engendering a level of distrust and sense of disenfranchisement among the population writ large—a failure of democratic legitimacy. In the process, the People lose effective control over the helm of national security policy; and, institutionally speaking, once lost, such control will take time and considerable effort for Congress to regain.

B. The Fallacy of Imperial Warmaking and the Reality of Coordinate Powersharing

Were Congress unquestionably subordinated by an Executive authorized to assert exclusive powers to engage troops in combat unilaterally, then any separation-of-powers concern stemming from military privatization would fall to the wayside: Regarding the
deployment of U.S. soldiers in zones of hostility, without any obligation to consult with Congress, let alone seek its approval, it would make no difference at least from this perspective if the Executive outsourced military responsibilities to private contractors. Although other constitutional and prudential harms would still ensue, the structure of powersharing between the elected branches would not be destabilized as a result of privatization. But despite actions and rhetoric suggesting that the Executive possesses unrivaled warmaking authority, the Constitution does not grant the president those exclusive powers, and hence in order to grasp the very real threat privatization poses to the equitable and prudent allocation of war powers, we must appreciate Congress’s important role in military affairs.

The exact contours of congressional-presidential powersharing need not be explored here; nor need we debate which branch, if either, has a preponderant say in decisions to commit troops. Those critically important questions are beyond this inquiry’s ken. The more modest aim is simply to establish the existence and persistence of a strong congressional role in military affairs both as a vital check on the Executive and as a necessary conduit to ensure the continued informed consent of the American people. In what follows directly below, I describe the principal ways in which Congress typically plays a prominent role in shaping military policy and, concomitantly, in constraining presidential unilateralism. Then, in Section C, I will discuss how privatization allows the Executive greater leave to bypass congressional oversight and authorization in those key domains. I do note at the outset, however, that congressional authority over the affairs of the U.S. Armed Forces is not perfect; nor is Congress entirely unable to oversee the activities of military contractors. Accordingly, though I do want to highlight the important differences between Congress’s influence over the Armed Forces and its influence over military contractors (both in theory and in practice), I recognize that at times these differences are ones of degree, rather than of kind.

Congress tends to exercise its authority over military policy along three key axes: its power to regulate military personnel, to appropriate funds to the military, and to authorize the deployment of U.S. combat troops in conflict zones. First, through its authority to *regulate military personnel*, Congress can constrain presidential warmaking by limiting the size of the U.S. military, by imposing restrictions and regulations on how and where soldiers can be deployed, and by structuring the chains of command to limit an Executive’s ability to politicize the military leadership.

Indeed, by possessing power over the conscription of American civilians and by regulating the standards of reserve activations, Congress can potentially limit the size of a conflict and its relative duration. Without the prospects of an unlimited, fresh supply of troops as replacements and reinforcements, the president may feel constrained in initiating and continuing unilateral engagements. Also, Congress can impose rules regarding the internal governance of the military, set terms for the conduct of war, and establish restrictive guidelines for engagement. In the absence of this set of Article I powers, the president—as Commander-in-Chief—presumably would possess the exclusive authority to determine the acceptable contours of soldierly conduct. And finally, still within this first set of powers, Congress can limit the politicization of the military by legislating hierarchical promotional guidelines and by organizing units around
civilian and military leaders whose positions require Senate confirmation pursuant to the Appointments Clause. In all, this first category of checks constrains the exercise of unbridled presidential warmaking and adds layers of transparency vis-à-vis fixed rules of military conduct and decisionmaking that ensure greater public awareness of military policymaking.

Second, another critically effective axis-of-constraints check on executive-driven military policy is Congress’s *power of the purse*, perhaps its ultimate trump card. Appropriations decisions, which belong to Congress and within the context of U.S. military spending *must* be constitutionally revisited at least every two years, are often “conceived of as lump-sum grants with ‘strings’ attached . . . binding the operating arm of government.” This power was notably employed in the Vietnam era, when Congress cut off all funds for use in operations in Cambodia; then, a decade later, Congress again tightened the purse strings to limit the president’s efforts in Nicaragua; and, in the present, post-Cold War era, Congress has used its appropriations power with some regularity to limit presidential power and narrow the scope of military engagements in Haiti, Somalia, the Balkans, and Rwanda. As will be discussed below, for a number of reasons it is much more administratively difficult to regulate the funding that ultimately flows to privateers, because contractors are often paid through more discreet, even convoluted bureaucratic channels (if they are even paid through the U.S. Treasury at all). Of course, it could be the case that Congress, in making its appropriations, begins with the foundational assumption that executive agencies—even ones not directly involved in national security—employ military contractors. But while this may very well become a commonplace assumption among future congresses, it is doubtful that previous and even the contemporary congresses, which appropriated war spending in Afghanistan and Iraq, contemplated the scope (and complexities) of private military funding.

Third, there is Congress’s most direct (but also most contested) power: to *authorize military deployments* even short of a formal declaration of war. Today, pursuant to the War Powers Resolution, a statutory rule ensuring that the collective judgment of both elected branches will apply to military intervention in a manner consistent with “the intent of the framers,” the president must consult with Congress and ultimately seek its approval to deploy and retain U.S. military forces in zones of combat. Despite opposition to this statutory framework and a refusal to concede that Congress has any role to play in military engagements short of formal war, recent administrations have nevertheless consulted with Congress—and sought formal authorization—*before* deploying troops for combat duty abroad. For example, both President George H.W. Bush in the Gulf War and President Clinton in the Balkans and Iraq aligned themselves with, rather than against, the powerful argument that Congress should take responsibility [in war decisions]. Each President submitted his defining military action to Congress along with a request for congressional approval in advance. Thus . . . highly publicized congressional debates characterize the present arena for resolving questions about putting the military in harm’s way.

And notwithstanding otherwise embracing the pretensions of vast executive prerogative, President George W. Bush has followed his predecessors’ deferential lead by seeking
congressional votes of support and authorization before taking up arms in both Afghanistan and Iraq.

C. Bypassing Congress Through Privatization: An Attack on Constitutional, Limited Government

1. Denial of Congress’s Regulatory Role

   a. Size of Military

   ... As mentioned above, Congress can preemptively constrain the excesses of a hawkish president by limiting the number of available troops. With a finite-sized public military, the president must deploy troops judiciously, or otherwise be forced to ask Congress to authorize a draft, liberalize reservist activation policies, or slowly expand through recruitment and retention programs. Any such request by the president to Congress would invite questions and criticisms of current strategies and priorities. The president’s expectation of political opposition provides crucial ex ante checks on executive adventurism and thus has the effect of counseling caution in how soldiers are deployed around the world. The other option for a president constrained by the size of the military is also disastrous politically: The overextended president (unwilling to request a draft) might be forced to withdraw troops from a conflict zone prematurely, and face the inevitable criticism for starting a war that could not be successfully completed.

   If on the other hand, there were some external, elastic source of troops, who could complement and supplement the U.S. Armed Forces, provide needed reinforcements, and help the president avoid having to activate reservists and/or reinstituting a military draft, the costs of not acting conservatively and judiciously are lowered. Privatization, at least at the margins, therefore presents a great alternative to lobbying Capitol Hill and the American people for permission to increase the size of the military quickly. While contractors could not “discreetly” command an entire theater in a major conflict, smaller outfits can be selectively positioned to provide the president with much greater flexibility—to continue, for instance, an unpopular or unexpectedly demanding war (that neither the president nor Congress would want to bolster with fresh newly conscripted soldiers). Hence with lower political opportunity costs for waging war, the president may be more apt to overcommit American capital—human, monetary, and diplomatic—in ways that would be less likely to occur were Congress and the American people (through their legislators) given a more direct say.

   One need not ponder hypotheticals to appreciate this potential for dangerous presidential unilateralism. If it were not for the tens of thousands of private troops supporting and serving alongside of U.S. soldiers in Iraq and Afghanistan, perhaps the President would not have been so eager to invade Iraq; or, perhaps, the limited number of American troops available would have compelled him to seek a broader coalition of countries willing to commit their own personnel to these endeavors at the outset. By relying on external, private sources for troops, the President has, perhaps, overextended American obligations abroad, turned his back on collective security measures, and in the process drawn the ire of a great many. (Hence, these “structural” harms are independent of any

Draft — do not distribute!
accountability-related transgressions that privateers might themselves perpetrate once deployed.)

Accordingly, tapping into an external, elastic supply of contract personnel could breach a tacit—and, no doubt, often hard fought—agreement between the Executive and Congress on the size of the military. This harm is, immediately, a fiscal one: it might be the case that Congress and the president agreed to keep the military comparatively small to reduce expenditures and reap peace dividends after, for example, the thawing of the very costly Cold War. But, the harm is also a political and legal one: Perhaps Congress kept the military small to dissuade an overly interventionist president from participating in far-flung engagements. Moreover, Congress might have agreed to authorize specific war powers requests only with the knowledge that the engagement would be of a limited scope commensurate with the manpower resources it assumed were available. Again, to the extent that the president could extend the duration and expand the magnitude of war by employing private contractors and to the extent that Congress had not been anticipating the wholesale reliance on military privateers, privatization provides opportunities to subvert these carefully arrived at arrangements.

**b. Reporting and Oversight**

Another key constraint on the president’s conduct of war takes the form of Congress’s reporting and oversight functions. Consultation with, written reports to, and oversight hearings before Congress represent important ways in which military policy is subject to considerable scrutiny and accountability. Typically, Congress has opportunities to debate and hold hearings on matters of national security—shedding light and imposing accountability on the Executive Branch. If any given deployment of forces would be received critically, say, as overly dangerous, destructive, or antithetical to American principles of democracy, an administration might be deterred from pursuing such ends in the first place. And, even if the Executive, wanting to avoid the use of actual soldiers (because of the reporting requirements under the War Powers Resolution) used CIA operatives, a framework of reporting and oversight statutes are in place to ensure a modicum of accountability and transparency over those individuals too. But when neither members of the U.S. Armed Forces nor other government officials are intimately involved in a particular engagement, it is quite possible that members of Congress would not be as fully informed about the activities being undertaken by private contractors.

Private firms thus permit the president to conduct military operations (especially small-scale ones not involving many, or any, U.S. troops) without having the same obligations to notify and involve Congress as would exist were American soldiers used. Privateers can, moreover, be contracted into service through third-party nations, host countries, or quasi-independent agencies, as has been the case with some American-based firms operating in the Balkans and even in Iraq. In these instances, Congress has comparatively little oversight authority. Indeed, the principal federal law, the Arms Export Control Act (“AECA”), which, inter alia, sets the terms by which information about American contractors working for foreign nations must be disclosed to Congress, currently requires the State Department to notify Congress only when a contract it authorizes exceeds $50 million.
And, even if the privateers were operating directly for the federal government, their contracts might (purposefully or unintentionally) have been indirectly routed through the Commerce, Interior, or the State Department, rather than the Defense Department. The congressional committees that principally oversee Commerce and Interior, for example, may not be sufficiently informed or interested, and may not have developed the requisite expertise to be effective monitors of such contracts. Finally, even if the contracts were issued through the Pentagon, matters of military privatization may not arise per se—short of a massive fiasco such as the Iraq prison-abuse scandal—that would warrant congressional interest from the Armed Services committees. This is, again, not to say Congress is unfailingly vigilant with respect to oversight of “public” military affairs, and entirely enfeebled with respect to overseeing military contractors. But while recognizing that the differences in congressional oversight are quantitative rather than qualitative, they are nevertheless important.

These oversight difficulties cannot be reduced to mere accountability lapses. Rather, these oversight difficulties also sound in terms of structural concerns about the architecture of American government. Even if Congress insisted on more centralization in the contracting process, because of the nature and design of military contracts and because of issues of private-sector proprietary information more generally, it is still questionable whether adequate information would readily be disclosed to an oversight committee were either a private military firm or a government official subpoenaed and asked to testify about critical details of an agreement. This proprietary information concern has already become a major source of executive-congressional tension in the commercial military contracting realm. One notable example involves the Air Force invoking the principle of proprietary information to fend off repeated requests by Congress to disclose certain information regarding its Tanker contract with Boeing.

Therefore, with limited congressional oversight and reporting, there are comparatively fewer political and legal checks constraining how the president conducts military affairs. The Executive’s policies may not be in line with the priorities and principles of Congress and the American people, such as when, for instance, the State Department, under the AECA framework, approved requests from MPRI to perform military consulting services for the repressive regime running Equitorial Guinea as well as for the Abacha dictatorship in Nigeria. It is at least debatable whether such permission would have been as readily granted were congressional consent a bona fide prerequisite. And, strategic interests and prudential policymaking aside, the lack of effective oversight deprives Congress and the People of an opportunity to debate normative concerns about delegating governmental policymaking decisions to privateers in the first place.

Accordingly, circumventing congressional oversight lengthens the leash the Executive has in conducting national security policy and, concomitantly, limits the effective transmission of information to the American public.

c. The Appointments Clause: Senate Confirmation of Military Officers

Since military officers are “appointed in the manner of principal officers [of the United States],” every individual, holding at least the rank of second lieutenant or ensign must be
nominated by the president and confirmed by the Senate. The Senate must also confirm the commissions of all reservists above the level of major. And, each time an officer is promoted to a higher rank, another round of Senate confirmations is required.

Though it is rare and administratively difficult for either the president or members of the Senate to be intimately involved in, say, the promotion of any particular Army captain, at the higher levels of military commissions, individual evaluations and considerations become more commonplace. In those cases, where appointments are important in shaping the policy direction as well as the public image of the American military, both presidential and Senate scrutiny is evident. Importantly, however, as Justice Souter noted in Weiss v. United States, many of the military officers subject to Senate confirmation are, constitutionally speaking, “inferior officers” that do not require the advice and consent of the upper house. But Congress has not chosen to vest the appointment of those (inferior) officers in the president and, instead, continues to subject those officers to the “rigors” of Senate confirmation; Congress’s decision not to abdicate this responsibility suggests that the Senate values and takes seriously its oversight role in this capacity.

If contractors carrying out American policy are not vetted through the process of presidential appointment and Senate confirmation, it is questionable whether, given the Senate’s oversight of military officers’ nominations down to the level of ensigns and lieutenants, they possess the legal authority or legitimacy to exercise the lethal discretion bestowed on them. As discussed, at Abu Ghraib, private contractors with little oversight were allegedly given broad (officer-like) discretion in helping set and implement interrogation policies and, in turn, were themselves issuing to U.S. enlisted soldiers orders that included the directives—ostensibly speaking—to brutalize or humiliate detainees. Whether bypassing the appointments process is a deliberate aspect of the decision to privatize or, more likely, an unintended consequence of the outsourcing objective, the fact remains that contracting out the responsibilities of active military engagement to ersatz “officers” deprives the Senate of one of its core duties—as applied both as a check on an injudicious Executive and as a safeguard for continued civilian control over the military.

Presumably even if the confirmation process is not treated with the individualized attention given, for example, to Supreme Court nominees, the Senate could still insist that all prospective nominees to command positions must satisfy certain blanket requirements. Those might include an absence of any type of criminal or domestic-abuse citation to ensure that the military advances only those individuals with impeccable professional and ethical credentials. Without such review processes, privatization (as in Abu Ghraib) may continue to permit the advancement of a range of less desirable candidates who lack the moral virtues and skills necessary to lead by deed and example.

Of course, since many military officers also were intimately involved in the prison-abuse scandal, clearly the appointments process alone is not a dispositive factor. So while I do not want to overstate the importance of the Appointments Clause, I do note that it would be significantly easier to conduct more searching review processes for military officers than it would for both the House and Senate to pass—and the president to sign—
comprehensive legislation regulating and, perhaps, licensing the types of employees that military contractors can hire.

d. Governance and Discipline of the Military

Finally, the Constitution authorizes Congress to establish codes of governance for members of the U.S. Armed Forces. Congress sets disciplinary guidelines for soldiers and authorizes the imposition of penalties in the event that they violate their oaths of duty or engage in any other form of proscribed behavior. Civilian contractors are not (and perhaps cannot be) effectively regulated to the same extent—and thus this status differential between contractors and soldiers may provide the Pentagon with opportunities to permit practices and behaviors (such as physical abuse for the purpose of extracting information) that are otherwise off-limits to U.S. troops. Leaving that insidious possibility aside, this issue of discipline via Congress is important because the Constitution separates the command of the military from the governance of the military, presumably to prevent an aggrandizement of war powers. But military discipline is broader than a separation-of-powers matter because the president, even as Commander-in-Chief, also may not be able to control contractors to a satisfactory extent. Part of this difficulty in disciplining contractors as if they were soldiers is that the Supreme Court has given Congress virtually plenary power to regulate the behavior of military personnel, and it is at least an open question whether the Court would also permit Congress to impose similarly strict rules backed by criminal punishments on top of—or in lieu of—contractual arrangements with privateers absent a formal declaration of war. Accordingly, because of its complexity and because it is not just a separation-of-powers concern, this subject will be treated at greater length and with broader sweep in Part IV.

2. Denial of the Appropriations Role

By using private contractors, the president may also reduce the likelihood of Congress easily terminating military funding. The sources of funds for private guards in Afghanistan, for coca-crop dusters in Colombia, and for security forces in Iraq may be outside the formal scope of Defense Department appropriations budgets, and hence may be buried within longer-term funding sources that are not as readily apparent to Congress. As noted above in the context of identifying oversight difficulties, when contracts with privateers are scattered throughout or among executive agencies, it becomes very difficult for Congress to detect, target, and—if need be—attack particular streams of funding in order to influence policy via the purse. Congress could, of course, always strike at the Pentagon’s budget writ large in lieu of trying to track down discrete funding sources to privateers, but the political fallout for not appearing to support America’s troops and war effort may be too great of a disincentive.

And perhaps most troubling from a legal-control vantage point, sometimes military operations are funded off-shore, by host countries or sympathetic third-parties. This was the case in Bosnia, where for a variety of reasons, a coalition of Muslim nations paid the American contractors for services rendered. Obviously, when engagements are financed from sources outside of the U.S. Treasury, Congress’s power of the purse may not be an effective constraint. This is also somewhat of the case in Iraq, where a percentage of the
funding for operations (including security operations) administered through the CPA supposedly came from Iraqi oil sales and thus was disconnected from the federal fisc. Hence from an appropriations standpoint, there may be occasions where Congress’s influence is quite weak. Therefore, without yet another check, Congress and the American people not only have fewer means of halting operations they deem to be counterproductive, but they also have a more limited appreciation of how well-funded select operations in general may actually be.

3. Denial of the Authorization Role

Regulating military personnel and patrolling funding allocations are secondary weapons in Congress’s quiver. The degree to which Congress can regulate personnel and require testimony and briefings may have a modest impact on fundamental presidential decisions to deploy and direct forces in zones of conflict. This is not to diminish the importance of these congressional powers, but rather to acknowledge their individual limitations in terms of influencing and altering executive policymaking. When aggregated, however, these powers loom larger: Congress’s cumulative ability to limit troop size and to curtail funding and to insist on oversight briefings weaves a thick web of checks possibly sufficient to constrain unilateral action (and more certainly sufficient to provide incentives for the Executive to want to work closely with Congress).

When we turn to the issue of express authorization, however, Congress’s power is immediately evident. While often insisting that congressional authorization is unnecessary, presidents—especially over the last decade—have routinely if begrudgingly sought congressional resolutions in support of military action. Hence the authorization power does serve as a considerable constraint...

Without the customary and statutory need for ex ante consultation and authorization, the president could deploy private troops in a way that otherwise would never have dared been initiated if limited to American troops and, correspondingly, beholden to the dictates of the War Powers Resolution. But since the War Powers Resolution applies only to the deployment of U.S. Armed Forces and, moreover, since anti-covert operations legislation requiring congressional notification and consultation applies only to members of the U.S. intelligence community, there is room to maneuver unilaterally if the president were to use privateers.

The drug war in Colombia provides an apt example. Due to frustrations associated with Congress’s stringent limitations on the number and responsibilities of American soldiers in Colombia in the 1990s, private military firms were utilized probably in no small part to circumvent these legislative restrictions.

Again, the structural damage is clear: through bypassing Congress—and the American people—the Executive can initiate more conflict than the public might otherwise have been willing to support. And, extending the War Powers Resolution to contractors—though possible (as will be discussed in the Conclusion)—would be politically very difficult given the troubles Congress faced trying to pass the 1973 legislation over the
President’s veto (and that was when antiwar sentiment and hostility toward presidential warmaking power were both exceedingly high).

D. Bypassing the People Through Privatization: Harms to Democracy

Having explored how privatization can short-circuit the effective workings of constitutional government as a government of checks and balances, I turn now to the corollary harm: how privatization, by bypassing Congress, can damage the proper functioning of democratic government as one predicated on informed, popular consent. To the extent privatization permits the Executive to carry out military policy unilaterally, without consulting Congress and without seeking formal authorization, it circumvents primary avenues through which the People are informed and blocks off primary channels (namely Congress) through which the People can register their approval or voice their misgivings.

In short, the legitimacy of military policymaking depends not just on broad congressional involvement, but also on democratic input and popular consent.

Thus, when, or even if, the public is potentially precluded from taking part in such discussions, the democratic integrity of the country is greatly compromised.

Privatization creates opacities that may occlude the ready awareness of events. Americans who are unwittingly kept ill-informed of their country’s involvement in matters overseas cannot serve their necessary roles in keeping the State responsive and responsible. Conversely, when they are made aware of such engagements, they can express opposition or consent, organize parades or protests, enlist in the military as a sign of support or burn draft cards as a sign of disapproval. However inconvenient it might be for the Executive to be constrained by the opinions of the People, public participation is a necessary and valued component of the republican system as evidenced in the Constitution, culture, and customs of the United States. To use privatization to limit public disclosures and curtail public debates is to diminish popular sovereignty. But even without that intent on the part of the Executive, privatization has the effect of circumscribing not only Congress’s deliberative role, but also its oversight role, and thus, it could end up limiting the information that reaches the People.

Congress’s constitutional role in preserving popular sovereignty is, of course, critical—and revealing. Far from simply serving as an institutional counterbalance to the president, the architecture of congressional responsibility in warmaking bespeaks an express recognition of the imperative to keep the public informed and to keep elected officials responsive. Just as it is endowed to do in the context of presidential treaty-making or ministerial appointments, the Senate, on its own, could have been exclusively entrusted to resist the tendencies of an imperial president bent on unilaterally sending troops into zones of hostility. At the Founding, however, Senators, like the president, were not directly elected by the People—only the House was. So if congressional warmaking authorities were vested only in the Senate (as Hamilton originally proposed), one might read the Constitution as saying that although the Executive must be kept in check by a competing branch of government, there is no corresponding responsibility to ensure that
the People (through their biannually elected representatives in the House) would be given a say. But, because the entire Congress was and is empowered in matters of authorizing and funding wars, evidently there is a compelling democratic element to the allocation of war powers that complements the limited-government analysis discussed above in Part III.C.

The expectation of popular ratification of war does not only follow from the fact that the Constitution gives the lower legislative house a role in decisionmaking; the Constitution provides additional support. Many believe, for example, that the Second Amendment embodies a popular-sovereignty right vis-à-vis military matters. . . .

Indeed, the Second Amendment gave the People a physical “say” over the conduct of war by limiting the capacity of the federal executive to aggrandize central military power. In providing for the dispersed ownership of weapons by the citizens, the Founders envisioned the existence of a people’s army, and thus vested decisions over matters of defense in the hands of people, and communities. Since, at least in the premodern era of warfare, weapons had to “be carried onto the field by persons, the leaders [had to] address the population and persuade them to carry those guns.” This understanding comports with Akhil Amar’s as well. Professor Amar understands the Second Amendment as originating out of Locke’s recognition that “the people’s right to alter or abolish tyrannous government invariably required a popular appeal to arms,” and as reflecting a deep anxiety about a centralized federal military.

Today, of course, the role of the militia (and the relevance, at least in this context, of the Second Amendment) has been diminished by the needs of a standing professional army. But that spirit of popular sovereignty has endured and surfaced elsewhere, often at the intersection of war and voting: “Apparently it takes war to open the eyes of America to the injustice she imparts to her young men. For it is surely unjust and discriminating to command men to sacrifice their lives for a decision which they had no part in making.” Hence, as early as the Revolutionary War, the franchise has been expanded and enlarged at times of combat to accommodate not only the service of soldiers for their patriotic labor, but also out of recognition for their desire to have a say over the conduct of the war. That tradition of expanding and protecting the franchise for soldiers has continued throughout the decades and centuries. President Lincoln insisted, for example, that the nation hold presidential elections in 1864, in the midst of the Civil War, even though he knew that his defeat would likely lead to the abandonment of efforts to preserve the Union. And during World War II, the United States passed

[t]he Soldier Voting Acts of 1942 and 1944 [that] not only guaranteed soldiers and sailors overseas the right to vote during World War II, but also served as an opening wedge in the battle for poll tax repeal and other congressional action to guarantee the voting rights of blacks more generally.

More recently, the democratic linkages to war have been exemplified by the Vietnam era’s constitutional amendment that lowered the voting age from twenty-one to eighteen and thus addressed the perceived injustice of denying young soldiers and draftees a formal voice in the direction of war efforts. These tangible connections between war and
democracy have prompted Professor Pam Karlan to assert that “virtually every major expansion in the right to vote was connected intimately to war.”

Accordingly, with a built-in expectation of involvement in matters of war, any effort deliberate or otherwise to bypass Congress—and concomitantly—the People is a direct blow to the vitality of America’s democratic system. The unauthorized wars in Laos and Cambodia during the Vietnam conflict and the covert operations to prop-up anti-Communist regimes throughout the 1970s and 1980s in the Americas led to great disillusionment and distrust. It is the People who have been assigned the constitutional right and responsibility to register or withhold their informed consent. Anything serving to undercut that right threatens the legitimacy of the government.

IV. UNDERMINING THE INSTITUTIONAL INTEGRITY AND STRATEGIC COMPETENCE OF THE U.S. MILITARY

Even if Congress, and the People, were broadly informed and consulted about the shift toward privateers—and even if privatization were explicitly authorized by Congress—serious structural harms could still flow from the delegations of military functions to the private sector. In this Part, accordingly, I describe how contracting for the services of private troops, either to serve alongside U.S. military personnel or to operate by themselves, engenders significant institutional harms, strategic liabilities, and morale problems. First, because the Uniform Code of Military Justice does not apply to privateers, there is a greater possibility that contractors would distort a mission or abandon it altogether. This harm transcends the mere accountability concerns that can be remedied through more stringent oversight and more careful contracting. Indeed, it is not so much the possibility that privateers will fail to carry out a mission that is the principal concern; rather, at issue is the weakening of military justice and discipline on the battlefield that could upset civil-military relations and delegitimize democratic warmaking. Accordingly, as I will discuss below, to ensure military contractors comport themselves with the same discipline and restraint expected of regular soldiers, absent a congressional declaration of war, constitutional reform (not simple legislation) might be required.

And, second, I also explore in this Part a concomitant harm: how privateers who participate in U.S. military operations might tarnish public perceptions of the American military and debase the iconography of soldiers as citizen-patriots. Indeed, placing contractors alongside (or in lieu) of soldiers may ultimately damage the privileged normative status the American military has historically enjoyed. This too is not readily remedied through accountability-oriented, or simple legislative reforms.

A. Harms to the Institutional Integrity—and Comparative Excellence—of the American Military

1. The Notion of “Separate Community”

Regardless of whether she is stationed in Tikrit or Fort Dix and whether she is rounding up POWs or walking her dog on the base, the American soldier—from private to four-
star general—lives in a “separate community.” Members of the U.S. Armed Forces
operate within a unique constitutional framework of governance and discipline necessary
to ensure that they serve as effective yet restrained actors in national defense. Simply
stated: the American people entrust to their soldiers the awesome tools of devastating
destruction, as well as an equally awesome democratic authority to wield them. In return,
the People insist that their delegates on the battlefield are rigidly disciplined and handle
their responsibilities with great humility and humanity. . . .

Therefore, since the military has a sacred duty to carry out the directives of civilian
authorities to a tee, it is crucial (not only for the success of missions, but moreover, for
the enduring legitimacy of democratic warmaking) that under no circumstances will an
order be ignored or distorted. This degree of absolute and uncompromising discipline
requires a constitutionally separate governing infrastructure, far stricter than ordinary
civil and criminal codes promulgated by civilian governments and necessarily entailing
some loss of the ordinary and even constitutional rights citizens of the United States
otherwise enjoy. In other words, “[a]n Army sent into combat by a democracy cannot act
like one.”

Accordingly, for generations, the American military community has been a social, legal,
and economic entity [unto] itself; systems have been in place—in one form or another—
since the dawning days of the American Revolution to treat members of the U.S. Armed
Forces differently (and more restrictively). In 1950, Congress introduced the modern
incarnation of this separate system: the Uniform Code of Military Justice (“UCMJ”). The
UCMJ represents an entirely endogenous value system that recognizes the weighty
authority and discretion given to soldiers and attempts to control that authority and
discretion more stringently than regular American constitutional and statutory law would
ever permit. The code subjects to military discipline, and at times to court-martial, those
individuals who are, inter alia, AWOL, disobedient, insubordinate, malingering,
misbehaving, or who render faulty performances of duty.

The UCMJ is more than a simple legislative enactment, but rather has the effective
currency of what Professors William Eskridge and John Ferejohn call “super-statutes”
and what Professor Gerhard Casper describes as a “framework statute” because it takes
on quasi-constitutional qualities and prescribes an entire positive code of regulations and
conduct, respectively. Indeed, the courts have recognized the special and distinct qualities
of this governing regime and defer to Congress even when the UCMJ limits soldiers’
constitutional liberties in ways unimaginable if ever applied to civilians. Given the
intrusive scope of the UCMJ, and the courts’ emphasis on Congress’s special Article I
powers over the Armed Forces qua Armed Forces, it may be unlikely that these
regulations could easily be extended and applied to civilians, even ones who serve as
privateers.

And, beyond the formal, legal structure of discipline, the military’s separate community
bespeaks a distinct social and moral experience. The cohesion of military units (and their
detachment from the outside, civilian sphere of life) creates camaraderie and engenders
an esprit de corps necessary for optimal performance on the battlefield—where it is said
that individuals put their lives on the line for one another as much as for their nation. This
inculcation of virtue and honor is accomplished through the “personal immersion” in the ongoing “collective narrative of [the] corps,” a narrative that is supplemented in part by an inward-looking sense of shared culture and in part by an outward-oriented aversion to what is perceived as the lax values of civilian life. Again, but for obviously different reasons than legal-constitutional ones, this esprit is difficult to extend to privateers via fiat—legislative or otherwise.

It is with this context and history in mind that the blithe introduction of civilian contractors into positions involving the exercise of sensitive military authority seems particularly dangerous and counterproductive—violating the carefully crafted arrangements established over time precisely to minimize the possibility that agents of combat will disobey their principals’ commands and/or abandon their comrades in the heat of battle.

2. Privatization’s Harms

Civilian contractors, not similarly subject to the dictates of military law or to the constitutional oath of office, cannot necessarily be expected or permitted to exercise the authority, judgment, or lethal force entrusted to soldiers. Contractors are not governed and disciplined by the same legal and socio-cultural obligations of duty and loyalty required to ensure the effective subordination of soldiers’ own interests and to guarantee the success of a given endeavor. No legal contract between the Pentagon and a private firm can hope to imitate, let alone replicate, this sacred relationship. Otherwise, why would U.S. military personnel be treated so differently than, say, civil servants working in the Department of Veterans Affairs? If American servicemen and women could be trusted to do their job as effectively without the UCMJ, then the entire legal and cultural architecture of the “separate community” would be largely unnecessary. The fact, however, that the separate community is so important to maintaining order and ensuring fidelity gives us a sense of why merely tightening contractual obligations and increasing contractor oversight might be all that would be needed when the government outsources commercial responsibilities at Veterans Affairs, but that those measures may not be enough when it comes to privatizing military functions.

Indeed, constitutionally speaking, it is at least questionable whether contractual penalties for violating many of the terms of a private military agreement can rise to the threat-level of an impending court-martial. Thus, given, for example, the Court’s historical jurisprudence invalidating laws that criminalize the mere breaking of private employment contracts, one might suppose that there would be some resistance to penalizing contractors as if they were U.S. soldiers (for all sorts of small infractions).

Since private agents are not controlled and disciplined by their governmental principals to the extent Congress requires and the Supreme Court allows for U.S. soldiers within the chain-of-command, it would seem inappropriate to delegate to private actors crucial military responsibilities, which require not only the careful exercise of life-and-death discretion, but also the internalization of civilian-military protocols regarding fidelity to officers’ orders. In short, contractors are not necessarily appropriately situated within the delicately woven legal and constitutional fabric that both endows the military with
authority to serve as an effective fighting force and, at the same time, severely curtails soldiers’ freedom to deviate in any way from their explicit charge.

   a. Potential Strategic Liability

First, privateers may at times prove ineffective, if not harmful. As already suggested, they are bound principally by contractual obligations to complete their missions—not by the command structure of the UCMJ nor, probably, by the ethos of honor and self-sacrifice cultivated within military units. Legal threats of punishment, or emotional appeals to fraternity or patriotism may not work to compel contractors to remain in harm’s way and accomplish their assigned tasks. Though these contractors may even be decorated veterans and steadfast patriots, no threats of courts-martial or fears that they will be harshly disciplined as deserters enter into their minds and oblige them to complete the assignments. . . .

Immune from the harsh measures of military justice intended to ensure no soldier will prioritize self-preservation over the good of the mission, it is more likely that key contractors, engaged in surveillance flights, responsible for caravanning necessary materiel to the frontlines, or defending key American installations in hostile territory, will simply shirk their duties. Moreover, among contractors there may not be the same psycho-social urgency to display true honor as a selfless contributor in the military effort.

The Pentagon is not unaware of the fact that when contractors are deployed, there is a greater likelihood of desertions and refusals to obey orders. As early as 1976, when tensions flared up on the Korean peninsula, a number of Defense Department civilian and contract personnel (rendering commercial services) made a mass exodus. Military officers could not “order” the contractors to stay and, as a result, their absence—to the extent their services were missed—compromised American and South Korean interests. More recently, the Pentagon commissioned a study that found commercial contractors might have fled the Persian Gulf theater during the first war against Iraq, were gunfire to have intensified or were Saddam Hussein to have unleashed chemical or biological weapons.

With this historical sensitivity to civilian desertions in mind, it seems somewhat reckless for the current Administration to have leveraged the battlefield and the post-war occupation with private contractors in Iraq—especially since this invasion was largely predicated on the U.S. government’s conviction that Saddam had (and was prepared to use) Weapons of Mass Destruction. . . .

Recently, contractors in Iraq have been put to the test and, by and large, have comported themselves quite admirably. Employees of Blackwater were besieged by insurgents and nevertheless ably defended an American installation without the assistance of U.S. troops. Obviously individuals who agree to serve as privateers in conflict zones are aware of the dangers, and companies and their employees who want to be repeat players have every incentive to exhibit that type of responsible, even heroic performance. Yet, on the aggregate, the possibility of desertions, acts of defiance, or reluctance to put one’s life on
the line is likely to be greater when individuals outside the special confines of the military community are delegated combat responsibilities.

b. Perceived Strategic Liability/Morale Problem

Moreover, even if the contractors do not appreciably undermine a campaign, regular U.S. troops’ misgivings may not subside—and for a justifiable reason: there’s always the threat that the contractors will walk out during the next siege. For the reason expressed above, the mere belief that contractors may flee is enough to introduce uncertainty and distrust among the U.S. troops—which is probably already high given the host of other existing morale problems currently plaguing the service ranks. And, the soldiers’ insecurity and their misgivings about privateers must be treated seriously; the military goes to such extensive lengths to engender the appropriate level of cohesion, discipline, and camaraderie that it seems inexplicable why the Pentagon would sacrifice those goals in the name of outsourcing.

An additional cause of concern from a morale and confidence-damaging perspective is the possibility that privateers will comport themselves in an unbecoming manner. Unhinged from the narrative of military honor, privateers may never have internalized the ethos of honor and dignity that is inculcated in American GIs. (And, even if the contractors are themselves veterans, that esprit may have long since diminished and been superseded by the mores of the marketplace.) As one recent observer of DynCorp’s behavior in Kabul noted, “[c]ontractors do not live by the same constraints as active-duty soldiers . . . [T]heir blurring of the military-civilian line serves as a reminder that military discipline not only keeps up morale, but encourages moral behavior.” American soldiers today (though admittedly not all model citizen-soldiers themselves) are taught the lessons of, for example, the My Lai massacre, and are told that those who helped stop the bloodshed were given medals[,] but that those who orchestrated it (and even those who just followed orders)[] were court-martialed. Situating soldiers in a storied tradition of honor may not eradicate all instances of criminal or excessively brutal behavior, but that educational process may inform the soldiers of the institutional condemnation that will be affixed to any such transgressions. It should not therefore be surprising that privateers, though hardly alone, were nevertheless at the center of the Abu Ghraib scandal in Iraq— involving the brutal torturing of Iraqi civilian prisoners—not just as participants, but as supervisors. Whereas courts-martial quickly followed for the U.S. soldiers involved, thus signaling (albeit belatedly) the government’s intolerance toward such behavior, it was reported that even after the news of the scandal broke and courts-martial were being convened, the contractors were still on the job, just as was the case with those DynCorp employees who ran a sex-slave operation in Bosnia. In the wake of that travesty in the Balkans, the only prophylactic measure taken by the company was to insist that each employee sign a statement saying she understands “human trafficking and prostitution are ‘immoral, unethical, and strongly prohibited.’” Recall, too, that DynCorp summarily fired rather than rewarded the whistleblower in that case. Since misdeeds like what happened at Abu Ghraib redound through the regular ranks of the military and lead to disillusionment and demoralization, the government, at least by staging investigations and courts-martial, can at least try to embrace a zero-tolerance policy and hope to rebuild
confidence among the rank and file and offer credible reassurances to Iraqis and the global community that such behavior is not condoned.

c. Perverse Incentives To Prolong/Expand War

An additional harm, which I discuss even though it may seem to be simply a conventional accountability concern, is the possibility that the incentive differential between soldiers and contractors could lead to mission distortions. Such an incentive differential (and the corresponding threat of policy distortions) is common, of course, in any number of other policy domains in which privatization has been introduced. Money after all is the reason contractors show up, and monetary considerations may skew the aims of the mission. Whereas presumably many regular soldiers would gladly forgo their “danger pay” to be stateside with their families and out of harm’s way, contractors’ livelihoods depend on the continuation—if not exacerbation—of conflict. Indeed, it is reported that military contractors have referred to the current administration’s reliance on military outsourcing as the “Iraq Gold Mine” and have likewise mused (quite presciently) that the fallout from September 11 would prove to be a privateer’s windfall.

Outfits paid a per diem may prefer to prolong the engagement, perhaps not working as swiftly or efficiently as they otherwise should. There have, for instance, been allegations that Halliburton has run additional but unnecessary supply convoys through Iraq because it gets paid by the trip. If true, this wasteful practice not only endangers the lives of Halliburton employees, but also U.S. troops, who may be dragged into the fray were an insurgent attack to occur. Hence, just as in other privatization contexts where monitoring is difficult or costly, private military contractors may deliberately take longer, say, to train and certify the competency of a domestic police force; or they may slow down their rate of coca-burning work to get paid for a few extra days or weeks. Alternatively, instead of sitting on their hands, they may have the converse—but no more acceptable—agenda: to be as destructive as possible. In this scenario, there may be an impulse to level rather than preserve since oftentimes it is the same (or related) firms providing security services in a zone of conflict that are also key players in physical reconstruction. A particularly devastated area may create the need for a government to issue contracts for road-building, public works projects, and security-training.

Again, this is not to say perverse incentives are unique to a military-contracting context. But because flying surveillance missions, destroying coca-fields, and providing security details abroad are not linear tasks that lend themselves to precise contractual regimentation and oversight, the agreements between the government and the firms must necessarily be somewhat open-ended; recognizing the uncertainties of dangerous assignments and crediting the service providers with the ability to adapt and change course when exigencies require doing so leaves the government vulnerable to more than economic abuses of the contractual relationship.

B. Debunking the Normative Iconography of the Citizen-Soldier

The introduction of private contractors—and their attempted integration into the American fighting forces—may also create a gap, a breach in America’s storied civic
republican narrative such that now, perhaps, military service to the State will be even more disassociated with notions of citizenship than it already has begun to be in this era of an all-volunteer military; indeed, taking up arms will be viewed even more widely as yet another commercial relationship, not totally unlike catering or maintaining public grounds.

Historically, Americans have looked to the moral authority of their country’s foreign policy and based it, on no small part, on the willingness of its citizens to put down their ploughshares and fight (and die) for a cause. To disaggregate that connection and commodify the role of a soldier as for-profit contractor may further separate the bounty of citizenship from the obligations that membership entails. That is, at a time when that connection is already tenuous—due in part to the replacement of a universally conscripted military with one comprised of volunteers—further disassociation through the practice of contracting out may prove quite disruptive.

1. First Among Equals: Traditional Laurels for Citizen-Soldiers

Many believe that military service is inextricably linked to citizenship, and vice versa. Accordingly, although this nation’s conception of service has changed over time, American soldiers and veterans have almost always enjoyed a preeminent status in our society. In a country of equals, founded on the rejection of titles, inherited or even merited, U.S. military officers are, perhaps uniquely, addressed by their command ranks long after their tenure in the military ends—a testament to their esteem in the eyes of the State and its citizenry as well as to the value of those titles above all others. True patriots from generals like Washington to grunts like Truman have taken up arms when their country has needed their service. And, like the ancient Cincinnatus, they returned home to civilian life when the fighting was done. This restraint, this willingness (if not eagerness) to beat their swords back into ploughshares and resume the business of ordinary living, has marked the American military as exceptional and amateurish in the noblest sense of that latter term.

Thus, in many circles, to be an American citizen is to be an American [soldier]. Anything short of that demonstrated commitment to the safety and security of the Republic relegates one to a lower rung of society, especially if one is an able-bodied male who intentionally avoided military service. Among other reasons, it is the reality of this socio-political hierarchy in America that makes the Pentagon’s refusal to permit gays into the military and to permit women into combat (not to mention a bitter history of discrimination against blacks) particularly painful and debilitating to those excluded.

I do not want to overstate this point as “soldier-worship,” especially in the post-Vietnam climate, when military service has lost some of its imperative and much of its status as an intuitive obligation of citizenship. But, throughout the longer history of this country (and perhaps increasingly again today), it is undeniable that the military “man”—be he a patriot-planter of the Eighteenth Century, a universal conscriptee in World War II or Korea, a draftee in Vietnam, or a volunteer from today—has been treated as a paragon of civic virtue for dutifully embracing this gravest of responsibilities of citizenship. For their patriotism and sacrifices, they have been duly rewarded, not through the currency of the
marketplace, but rather through the currency of the polis. Hence, in addition to the honorific awards associated with rank and merit, military service has also been directly rewarded with the wholesale expansion of the franchise, the furthering of civil rights, and the development of a welfare system that in part predates and in some instances has even outclassed that afforded to America’s widows and children.

In short, the normative treatment of the soldier is as a public figure and hero. Victorious or slain, he (and increasingly she, too) is one of our “boys” (or “girls”) in the field. The military of today may no longer be comprised of citizen-militias, but the resonance of its members’ public service cannot be ignored. Empirically speaking, it is of course still unclear how privateers will affect this conception of citizen-soldiers; but one might presume, as a starting point, that profit-seeking contractors would diminish the normative standing of soldiers in general.

2. The Marketplace Debases the Polis

We do know (or rather, we think we know) this: Contractors are not per se battling for their country and their countrymen. They are not fighting to defend some ideal, vindicate some set of rights, or achieve national honor. As mentioned earlier, they are not even looking to lay down their weapons and go home. Instead, when objectives are achieved, privateers almost by definition look forward to the next lucrative engagement. Hence, to transform and possibly dilute the public service of national defense by introducing profit-motivated contractors may very well debase and commodify what has been the highest civic calling this or any other republic has known.

Whether it actually does so or not, privatization appears to weaken this connection between soldier and citizen—a connection that might, as suggested above, already be tenuous in a military era characterized by an all-volunteer fighting force, which includes many who enlist, at least in part, for financial reasons. Simply stated, the outbreak of war constitutes an economic windfall for contractors. With this profit-motive comes a perversion: As Colonel Thomas Dempsey has put it, when an American soldier kills, it is “because [his] president told [him] to”. . . . If a contractor shoots someone, it’s for another reason: “to get paid.”” This distinction, though perhaps overstated here, may not be lost on the American people, especially given the tenor of the news coming out of Iraq in 2004. During the same week that Pat Tillman, a former NFL standout died in Afghanistan, news broke of the central role privateers had in abusing Iraqi prisoners. The contrast between an All-American gridiron hero who gave up millions of dollars and his prime years as a professional athlete to enlist in the Army and an unscrupulous contractor brutalizing Iraqi detainees could not be starker.

The damage here could run beyond morale concerns just on the frontlines; Americans’ pride in their “boys and girls” may be dampened not just by the dismay felt at the appalling acts of brutality perpetrated under the American flag (by soldiers and contractors alike), but also by what they may view as the commodification of war, killing for money. In large part, the laurels bestowed on soldiers are premised on their endangering their lives to promote an ideal, preserve justice, or introduce freedom. Even knowing that, for many, their service is economically driven, i.e., performed with an eye

Draft — do not distribute!
toward learning trade skills or seeking the military’s help to pay for further education, we still consider today’s soldiers to be citizen-patriots. But we do not as readily reconcile economic self-interest and public service when it comes to contractors. If soldiers serve as liberators while incurring great personal sacrifices, then they are heroic; if they instead do so for profit, then they might tarnish the entire enterprise. Moreover, if the country’s pride and respect for its military wanes, perhaps more than the stature of GIs will fade. Perhaps veterans’ benefits will be cut—justified by the somewhat cynical expectation that those who serve in the armed forces are likely to find gainful employment afterwards as military contractors (even if many veterans consider the idea of private combat anathema).

Thus, strangely, a public disillusioned by military privatization might end up forcing citizen-soldiers into post-military careers as privateers. Even more significantly, American support for idealistic military endeavors, if those endeavors are perceived as being “corrupted” by the presence of profiteers, might similarly wane—and the notion of American intervention, humanitarian or otherwise, would then become less popular domestically as the public takes stock of who is fighting and for what reasons.

Or, returning to a point made earlier in this Article, a transition toward using greater numbers of private troops might engender the opposite result—more indifference to casualties and the lowering of the public’s apprehensions about waging war. If those fighting are not America’s boys and girls risking their lives to defend core interests of the State, but rather agents who voluntarily contract to perform explicitly dangerous missions, maybe the libertarianism of the market will outweigh the paternalism felt toward America’s soldiers and thus ultimately lower inhibitions against armed conflict. But in either case—if Americans are disillusioned with military action as illegitimate or if they become desensitized to combat losses—privateers could tarnish the luster of American foreign policy and spark immoderate feelings that may not match the goals and values of balanced, reasoned foreign policy. Again, tighter regulations or other reforms aimed at curtailing contractor discretion and contractor mismanagement are not, by themselves, capable of addressing these broad symbolic and cultural incidents of military privatization.

V. INTERNATIONAL LAW/DIPLOMACY HARMs

Having canvassed the constitutional, legal, and democratic harms in Parts III and IV, I turn now to the international/diplomatic harms privatization may cause. These harms pose considerable consequences for American foreign policy, for American credibility abroad, and for the interests of containing the proliferation of even less well-regulated military profiteering practices around the world.

A. Alienating Friends and Foes Alike

Contracting out allows the U.S. government to purchase strategic outcomes at a much lower political cost than if the boys and girls of America’s volunteer army were dispatched. Indeed, an overseas engagement involving contractors might, accordingly, produce neither an official body count nor much political opposition. But, the security
and flexibility the United States gains without expending domestic political capital and/or the lives of servicemen and women may, however, serve to validate the perception that the American agenda is driven by dollars rather than ideals; that decisions are made in private, smoke-filled backrooms rather than openly on the floors of Congress. It also invites concerns that the United States is represented in zones of hostilities by individuals who are not subject to the same standards of legal conduct and ethical restraint that this nation and the international community expects of the U.S. Armed Forces.

1. Allies

Among America’s allies, when the private cavalry is dispatched instead of the U.S. military, they may think that their particular crisis is outside of core American interests. This suspicion or sense of being slighted can breed resentment and a weakening of ties, a response not altogether lost on American leaders. Congressmen Tom Lantos and Henry Hyde had this precise concern in mind when they questioned the wisdom of contracting out President Karzai’s security detail. In a joint statement, they noted: “[T]he presence of commercial vendors [protecting Karzai] would send a message to the Afghan people and to President Karzai’s adversaries that we are not serious enough about our commitment to Afghanistan to dispatch U.S. personnel.”

Other allies too may be dissatisfied by the conduct of military engagements by private troops. No doubt the Bosnians would have preferred to receive the help of DynCorp contractors, without their extracurricular involvement in sex-trafficking operations. Moreover, perhaps pro-American leaders in the Middle East similarly feel betrayed, today, by the conduct of American privateers toward Iraqi prisoners. Leaders who endorse American foreign policy aims, often at great domestic peril, are then placed in an even more difficult situation at home when forced to defend their support in the face of American acts of brutality. Of course, transgressions by American soldiers certainly do occur. But, at least those acts can be reported up the chain of command and, in turn, can be swiftly punished, thus demonstrating the U.S. government’s commitment to justice and self-restraint; as we have discussed, comparable firmness with contractors is much more difficult to achieve.

2. Would-Be Allies

Let us also not forget that American military personnel are, increasingly, serving as diplomats, humanitarian providers, political consultants, and “liberators.” Their conduct on such missions could leave as large of an impression on their hosts as would any tangible project or aid package they deliver. Therefore, if the United States is dispatching private actors, who are not comporting themselves well, the conduct of these privateers will inevitably be imputed to all soldiers, if not all Americans, and the goods and services they provide will be, in the long run, devalued. As P.W. Singer notes, a “key realization of contracting is that a firm becomes an extension of government policy and, when operating in foreign lands, its diplomat on the ground. As such, the firm’s reputation can . . . implicate the government[’s] as well.”
And, finally, America acts not just as an intervenor or liberator, but also as an occupier. While on the ground, in Kabul or Baghdad, the U.S. personnel must work to win the hearts and minds of the locals. If American contractors were to act in an undignified, or offensive manner, it would only hamper the process of gaining the trust of the people. (Again, this assumes that because of the UCMJ and because of the military’s ethos of honor, soldiers are less likely to act inappropriately.)

3. Adversaries

And, among those who already consider America a corrupting force in the world, the privatization of military might, especially in efforts to circumvent U.N. agreements and arms embargoes, only further fan the flames of international dissent and discontent. The maniacal bombers of September 11 undertook diabolical deeds purportedly in the name of the disgruntled who viewed the World Trade Center and the Pentagon as the West’s twin evil exports. Amalgamating and conflating those formerly distinct entities via privatized war makes it that much harder to disabuse the world of its perceptions of the United States as an evil economic-military imperialist.

B. Flaunting the Ideals and Undermining the Institutions of Collective Security and Global Governance

The U.N. Security Council is widely viewed as the principal venue for deliberating on matters of collective security. Though hamstrung by internecine fighting among the permanent members during most of the Cold War, the Security Council emerged as an authoritative and relatively effective body in the early 1990s, serving as the centerpiece of what the first President Bush dubbed the “New World Order.”

For the most part, this renewed faith in the Security Council has been affirmed by member nations; but not entirely. Facing opposition on a proposal to intervene in Kosovo in 1999 and again, in 2002-03, on a decision to invade Iraq, the United States has forsaken the imprimatur of the Security Council and sought legitimation elsewhere. For Kosovo, America secured NATO’s approval; and for Iraq, the United States cobbled together a band of allies, euphemistically called the “Coalition of the Willing.” In the process of circumventing the United Nations, however, the United States has damaged the Security Council’s authority and called into question the credibility of collective security writ large.

Privatization only makes bypassing the U.N. easier and even more insidious than patching together an alternative source of collective authorization. At least with respect to small-scale interventions, where private troops could act in lieu of public soldiers, the United States could nominally remain a good global citizen and nominally recognize the supremacy of the Security Council, while still achieving those desired aims that the Council refuses to endorse. This would allow the United States to avoid the political backlash it felt (vis-à-vis Kosovo and especially Iraq) when it publicly eschewed the Security Council in favor of a more compliant authorizing community. For instance, say the United States or another member proposes a resolution in support of intervening in a small country, perhaps besieged by a humanitarian crisis or laboring under civil war.
Such a resolution fails. The United States can abide by the decision not to intervene formally, yet can still make available to the country in question a private American outfit to carry out the objectives that the Council rejected.

While this avenue of clandestine circumvention is, probably, unavailable in most instances where an effective force would have to be quite large, there are still opportunities in certain situations where small, discrete units would suffice. For example, small forces might prove especially useful in the nascency of attempted coups or during the early stages of civil unrest in the likes of Liberia, Sierra Leone, Sudan, or even Rwanda, where experts have now suggested that if intervention had occurred early enough, a crack outfit could have helped prevent genocidal civil war without the need for an overwhelming show of force. With the use of contractors, therefore, the U.S. government could also achieve some of its foreign policy ends, while not taking any responsibility for promoting them.

But the problem with contracting to avoid a Security Council veto is bigger than the mere issue of avoiding responsibility in any particular engagement: What is worse is that the nation would be turning its back on the legitimate collective security apparatus it helped found and promote, and would not even be doing so in a transparent way, i.e., calling for reforms to the Council’s procedures and operations or publicly shaming obstinate members. It would be more honest and responsible for the United States, if it were dissatisfied with some aspect of the Security Council, to seek direct reform. Such reform efforts would demonstrate the United States’s faith in the system of collective security and international law. But, to continue to operate outside its bounds, either via makeshift coalitions or private operations, while still purporting to respect the institution is to make a mockery of the Security Council and, moreover, to jeopardize the integrity of America’s foreign policy.

C. Setting Bad Precedents and Encouraging the Global Growth in Private Military Forces and Capabilities

Compared with foreign mercenaries operating elsewhere around the globe, U.S.-based privateers are relatively restrained. To satisfy both the generals in the Pentagon and the investors on Wall Street, American private military firms maintain a level of professionalism and decorum not always shared by their counterparts operating in other regions of the world. According to those who have surveyed privateers from a comparative perspective, there are major military firms based overseas, that lack the professional scruples that American companies appear to possess; simply stated, those firms are more likely to work for despotic or repressive regimes.

For example, major international military firms such as Executive Outcomes, Stablico, and Omega Support have each worked at various times on both sides of the Zaire-Congo conflict in the late 1990s. Executive Outcomes also helped the Sierra Leone government fend off rebel advances in 1995-96, and then had a hand in appointing an interim head of government—one reportedly with whom the South African-based firm could “work.” Evidence also points to the fact that Executive Outcomes considered the possibility of assisting the Rwandan Hutu government in 1994—not too far in advance of the time that
the Hutus were planning to unleash their murderous campaign against the Tutsis—and that Sandline came similarly close to working for the Mobutu regime in Zaire, despite its widespread notoriety as repressive and corrupt. More recently, a failed military coup in Equatorial Guinea involved privateers financed by, among others, the son of Margaret Thatcher. The goal, apparently, was to install a more business-friendly leader as head of the oil-rich state.

Private military firms help prop up rogue regimes, resist struggles for self-determination, and contribute to the proliferation and diffusion of weaponry and soldiers around the world—axiomatically a destabilizing and thus undesirable phenomenon. The existence of armaments held by stateless groups complicates the task for responsible countries who (for purposes of self-defense and collective security) keep track of and seek to contain the spread of weapons. The availability and acceptability of contractors makes it more difficult for countries to assess the relative strengths of rival nations, since one phone call to a group of out-of-work Ukrainian fighter pilots could radically alter a region’s balance of power. Of course, the existence of one such outfit also spawns greater demand—as every government would like the security of a few Ukrainian fighter pilots on retainer. Moreover, to the extent that privateers, especially those operating in Africa, may frequently be foreign nationals, the political and human costs of war may be quite low.

All of these factors point toward dangerous forms of military proliferation and thus threaten peace and stability. By all accounts, this global trend should be one the United States vociferously condemns. But can it do so credibly with thousands of its own privateers under contract? Even if the United States were to draw distinctions and make exceptions for its “professional” contractors, it probably still would be unable to lead a campaign against privateers. Therefore, privatization by the United States helps set a bad, enduring precedent and lends the global practice an unwarranted veneer of legitimacy.

VI. CONCLUSION

Given my analysis in the preceding parts, I might be tempted to conclude with the utmost of economy—and use only three letters: Q.E.D. However, despite the litany of structural (not to mention accountability-based) harms that private contractors introduce onto the national security landscape, it is doubtful that this new phenomenon—however much decried—will quickly fade away. Indeed, the combination of America’s extensive overseas military commitments, its already taxed store of reservists, and its inability to stomach a universal draft will probably ensure the continued need for an elastic supply of private troops for the foreseeable future. Hence, although this inquiry would rather conclude by way of proscription than prescription, realism preaches the latter approach might prove more prudent.

Appreciating the staying power of military privatization, critics and apologists alike have started to propose reforms centered on greater contractual control and oversight. These measures should, of course, be applauded as steps in the right direction. But, on their own, these reforms do not penetrate deeply enough to reduce the battery of structural harms military privatization creates; in other words, they do not go beyond accountability. Effective reform to combat the harms identified in this Article must
accordingly look beyond tightening contracting law and enforcing accountability norms. Reform instead must attack the underlying status discrepancies that distinguish contractors from U.S. troops. If there were no gaps in the way troops and privateers were governed, disciplined, and publicly perceived, then many of the institutional, legal, and symbolic distortions that threaten the vitality of America’s democratic institutions, the integrity of its fighting forces, and the legitimacy of its international relations would be greatly reduced.

Accordingly, by way of conclusion, I offer just a few words in service of sketching out a blueprint for reform. Irrespective of the particular details, any such blueprint to alleviate the structural problems raised in this Article must promote symmetry and parity between contractors and American troops in three distinct ways. First, reform must give Congress the regulatory and warmaking authority over privateers that is commensurate with what it enjoys over the U.S. Armed Forces. Achieving parity here, of course, would help minimize the democratic and constitutional harms that may exist today in situations where the legal status differentials between private contractors and U.S. troops can be exploited to circumvent Congress and the American people. Second, reform must enable the federal government to exercise the same amount of control and discipline over privateers as it does over members of the Armed Forces. Eliminating this disparity would help alleviate the concern that private troops representing American interests are—not effectively constrained by civilian and political officials. And, third, reform must somehow serve to lessen the symbolic differences between contractors and soldiers, differences that currently can be leveraged to deploy privateers in zones of hostility where the American people would not as readily commit public soldiers.

The crucial questions this reform agenda ultimately invites, then, are two-fold. Can these status disparities actually be eliminated? And, if so, does achieving parity reduce, if not altogether destroy, military privatization’s raison d’etre?

A. Achieving Parity: Leveling the Asymmetries Between the Public and Private

1. Restoring Equilibria in the National Security Constitution

For Congress to establish similar levels of control over privateers to that which it possesses over public troops, it must draft a comprehensive framework statute. Such a statute would both assert Congress’s authority over privateers and formalize the processes by which the national legislature is kept informed of their activities. This statute, for instance, might create a department within a federal agency, e.g., the Military Privatization Office of the Department of Defense (“MPO”), that Congress designates as the focal point for all military contracts and that has one assistant secretary in charge. The MPO would be guided by a set of clear regulations, one of which would necessarily be: no federal money can be disbursed to any contractor whose employees carry or fire guns overseas unless (1) the corresponding contract is routed through the MPO and (2) the recipient registers (like a lobbyist would) and discloses information about its employees, its clients, and its engagements. The MPO would be required to share this data with Congress and the public, as well as to report more generally—on, say, a quarterly basis—on the location, number, and activities of all contractors serving abroad as part of federal
contract work. Moreover, the assistant secretary would have to advise Congress within 48 hours if any of the contractors engage in gunfights and/or if there are any contractor casualties abroad.

Concomitantly, the statute would formalize a process by which Congress can officially authorize privateers to participate in conflicts. This could be achieved simply by extending the application of the existing War Powers Resolution beyond members of the Armed Forces to include contractors. Such a measure would work both to ensure congressional authority over privateers commensurate with the powers it possesses over U.S. military personnel and to increase the level of public awareness and transparency required not only for prudential, accountability reasons, but also to comport with the normative imperatives of democratic governance and popular sovereignty. Obviously, such legislation would probably be imperiled by the threat of a presidential veto and, perhaps, even a constitutional challenge (at least with respect to congressional insistence on its power to authorize contractor-led military engagements).

2. Disciplining Contractors as Soldiers

The next step would be for Congress to reduce the status disparities between U.S. troops and private contractors in the context of military discipline and control. Currently, these disparities generate reliability and dependability gaps: the government cannot impose the requisite discipline and penalties on privateers to ensure that they do not deviate from, undermine, or otherwise jeopardize a military mission through acts of insubordination and desertion. Such conduct disparities also spawn broader, structural concerns for the effective control and subordination of the U.S. military to the civilian federal government. As discussed at length in Part IV, whereas the UCMJ creates for U.S. troops an entire framework of discipline, *under pains of severe punishment*, there are no comparable disincentives that exist in the realm of contract law. The threat of a civil breach, even if its consequences ever run to the breaching privateer (and not just to the firm), cannot compare to the threat of being thrown in the brig.

Congress, of course, can extend the jurisdictional reach of the codes of American criminal law to contractors overseas—and has already done so with the Military Extraterritorial Jurisdiction Act of 2000. But for a variety of constitutional reasons, it may have more difficulty extending some of the unique rights-infringing provisions of the UCMJ to private contractors, or to any other civilians for that matter (at least in the absence of a concomitant congressional declaration of war). I have in mind here those behaviors or practices such as insubordination, criticism of military policy, and desertion that alone are not criminal acts (and arguably could not, constitutionally speaking, be *criminalized*) if applied to civilians, but which the Supreme Court has allowed Congress, in essence, to criminalize vis-à-vis active members of the Armed Forces precisely because of the military’s special, subordinated positioning in the architecture of American governance.

To narrow this gap, therefore, would probably require more dramatic measures than could be achieved through ordinary legislation. Perhaps, a constitutional amendment limiting the rights of overseas military contractors, an explicit step to “deputize”
contractors and incorporate them into the larger fold of the American military community, or a congressional declaration of war, which definitely extends military law to contractors working overseas with members of the Armed Forces, would be required to achieve effective control on par with what currently disciplines U.S. military personnel.

3. Cultural Conflation: Publicization of Contractors

The final step, then, would be to smooth out the symbolic differences between public and private troops. This would require (1) instilling in contractors a sense of their public charge (what Professor Jody Freeman has called “publicization”) and (2) conveying to the American people the sense that soldiers and contractors, symbolically speaking, are one and the same. Accomplishing these twin aims might increase morale among public troops worried that their private compatriots will not perform ably and reduce the status disparities that currently allow the president to overcommit forces for a given engagement, respectively.

In some ways, of the three sets of reforms, this is the most difficult task, because it is quite difficult for policymakers to legislate a change in perceptions (either among the contractors or among the American public more generally). It would be pointless, or at least inefficacious, to regulate how people go about valuing one life compared to another when any such exercise is inherently subjective and, likely, also idiosyncratic. But, on the other hand, perhaps the “publicization” follows closely—and somewhat effortlessly—on the heels of the foregoing tangible reforms. Perhaps with the incorporation of private soldiers into the regulatory framework already established for the Armed Forces, with the combining and conflating of soldier and contractor casualty counts, with comparable requirements of oversight and formal authorization, and, moreover, with some measures taken to discipline contractors like soldiers, contractors could feel more closely aligned with the American military and embrace its esprit; and the public, in turn, might begin to view contractors as more closely integrated into the American military community. Indeed, it may be the case that the symbolic differences are largely a function of status differences, and to the extent privateers are limited in their ability to act (comparatively speaking) ultra vires, they may not see themselves—or be perceived—as all that distinct from members of the Armed Forces.

Of course, some residual disparities of no small import will remain. For example, the social and emotional bonds forged while training and serving in military units, emphasized above as instrumental in fostering selflessness, valor, and an esprit de corps, cannot be transmitted to privateers just because they would now be governed by the same disciplinary standards and counted as soldiers for purposes of calculating force projections and body counts. Nor can the public be immediately persuaded that contractors and soldiers are one and the same just because they are similarly regulated. Ultimately, therefore, additional affirmative steps to integrate the two distinct cultures, such as by having privateers eat, train, and live with soldiers, by minimizing salary differentials, and by circumscribing opportunities for privateers to serve foreign clients ([thereby] minimizing reasons for the public to perceive them as mercenaries) would be necessary to reduce further the symbolic differences in a more meaningful way.
B. Coming Full Circle: Arriving at a Place Where Issues of Accountability and Efficiency Are (Again) Paramount

Assuming *arguendo* that the difficulties associated with devising and implementing the reform measures are overcome—and a comprehensive set of prescriptions do help eliminate many of the structural distortions that currently correspond with military privatization initiatives—there is an additional concern: Would closing these structural gaps destroy military privatization’s raison d’etre? In other words, does military outsourcing exist principally to leverage status differentials?

Closing the status gaps would indeed add to the “publicization” of private contractors. As mentioned above, if regulated and disciplined like U.S. troops and if the American people start thinking of them as comparable to U.S. troops, contractors may not be used as readily (or successfully) to exploit legal and symbolic asymmetries. This means, however, that policymakers could not rely on them to accomplish military objectives otherwise difficult to obtain, if not unobtainable, using U.S. soldiers. Yet room would still exist for private actors on the national security landscape: The economic-efficiency virtues of privatization would largely remain unaffected by the structural sets of policy reforms. Indeed, it is possible that contractors from the private sector could still offer the Pentagon high-quality services and lower prices. They could also provide the Defense Department with force-multiplying and specialization capabilities if additional troops are needed.

Arriving at that point, where the principal reasons for privatization center on economic efficiency gains, would, actually, permit scholarly analysis to come full circle as well. Once military privatization is stripped of its potential to be structurally damaging, it could then be scrutinized principally on accountability and efficiency grounds. That is, once the legal, constitutional, and symbolic concerns are allayed, we can be in a position to evaluate the true economic virtues of privatization (and the “inherently governmental” tradeoff), an inquiry I expressly bracketed for the purposes of this Article. It is then—and perhaps only then—that conventional discussions centering on costs and benefits, transparency, and accountability (all of which are very important) should resume in earnest.