2012

Rational Reasonableness: Toward a Positive Theory of Public Reason

Gillian K Hadfield, *University of Southern California Law*
Stephen Macedo, *Princeton University*

Rational Reasonableness: Toward a Positive Theory of Public Reason

Gillian K. Hadfield[*]
Stephen Macedo[**]

Why is it important for people to agree on and articulate shared reasons for just laws, rather than whatever reasons they personally find compelling? What, if any, practical role does public reason play in liberal democratic politics? We argue that the practical role of public reason can be better appreciated by examining the confluence of normative and positive political theory; the former represented here by liberal social contract theory of John Rawls and others, and the latter by rational choice or game theory. Citizens in a diverse society face a practical as well as a moral problem. How can they have confidence that others will reciprocate their commitment to supporting governing principles that depart from their own ideal conceptions of truth and value in order to be reasonable to all? Citizens face a practical problem of mutual assurance that public reason helps them solve, and solve as a matter of common knowledge. The solution, on both views, requires citizens' reciprocal commitment to basing law on a system of shared reasons. Both views place public reason at the core of liberal democratic politics in conditions of diversity, and for quite similar reasons. Our argument illustrates the (often) complementary roles of positive and values-based analysis in constitutional design.

---

[*] USC, University of Southern California.
[**] Princeton University.

For comments on earlier versions we are grateful to Nir Eyal, Dan Kahan, Rob Katz, Leif Wenar and also to fellow members of a discussion group at the Center for Advanced Study in the Behavioral Sciences at Stanford University: Victoria McGeer, Philip Pettit, and Daniel Posner.

Introduction

A. The Practical Challenge

The liberal idea of public reason generates tremendous controversy, which is echoed in much actual political debate. Many members of the U.S. Supreme Court and other public officials frequently seem to agree that "laws touching on constitutional basics ought to be supported by reasons that all citizens can share." Religious reasons, about which citizens reasonably and deeply disagree, cannot be the basis for a reciprocal commitment to living under laws justified and acceptable to all.

In response, religious citizens and their advocates, including many political and legal theorists, argue that the idea of public reason is gratuitously censorious. It inhibits some citizens' expression of their basic values on the most important political questions, provokes unnecessary conflict, and robs public debate of possible insight.[1]

Perhaps most surprisingly, it is often not clear what is at stake in this controversy. Public reason is often defended as part of "ideal theory": a feature of what John Rawls calls a "well-ordered society." But such ideals are remote, so why not focus on encouraging support for just laws irrespective of people's reasons? Why is it important for people to agree on and articulate shared reasons for just laws, rather than whatever reasons they personally find compelling? What, if any, practical role does public reason play in liberal democratic politics?

In this Article, we suggest that the practical role of public reason can be better appreciated by examining the confluence of normative[2] and positive political theory: i.e., distinct streams of research representing methodological approaches that are often treated as antithetical. The normative approach is represented here by a version of Rawlsian contractualism, which defends public reason as a duty of citizenship in a diverse constitutional democracy. With respect to the

---

[1] *See, e.g.,* Jeremy J. Waldron*, Religious Contributions in Public Deliberation,* 30 San Diego L. Rev. 817, 824 (1993); Jeremy J. Waldron, *Public Reason and 'Justification' in the Courtroom,*1 J.L. Phil. & Culture 107 (2007); Nicholas Wolterstorff, *The Role of Religion in Decision and Discussion of Political Issues, in* Robert Audi & Nicholas Wolterstorff, Religion in the Public Square: The Place of Religious Convictions in Political Debate (1997); Christopher J. Eberle, Religious Conviction in Liberal Politics (2002); Jeffrey L. Stout, Democracy and Tradition (2004); Gerald Gaus, The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World (2011).

[2] We recognize that "normative" and "moral" are not synonymous. Normative systems generate standards for assessment and reasons for action that need not be moral. A system of public morality is, however, a socially normative system, and public recognition of its socially normative character is crucial to its success. *See infra* Section IVA.

normative model of public reason, we do not simply channel Rawls but offer a critical interpretation and (in some respects) revision which Macedo has defended elsewhere[3]; the aim here is to advance the most powerful view, not Rawls. By positive political theory we mean rational choice or game theory, a framework of analysis based on the simplifying assumption that individuals are economically rational. This latter approach is represented here by a model developed by Gillian Hadfield and Barry Weingast.[4] These two quite distinct approaches converge on strikingly similar accounts of public reason.

We agree with Robert D. Putnam that truth is most likely to be found at the confluence of different streams of research.[5] An examination of the structural similarities of these differing modes of analysis will help explain why public reason is of practical importance to democratic politics. Our argument also seeks to illustrate what we believe is the (often) complementary role of positive and values-based analysis in constitutional (in the broadest sense) design.[6]

The two frameworks of analysis that we deploy here, and set in mutual dialogue, share a common problematic. Citizens in pluralistic societies—who by assumption do not all share a single conception of the good—face a practical as well as a moral problem. How can they have confidence that others will reciprocate their commitment to supporting governing principles that depart from their own ideal conceptions of truth and value in order to be reasonable to all? Citizens face a practical problem of mutual assurance that public reason helps them solve.[7] The solution, by both views, requires citizens' reciprocal commitment to basing law on a system of shared reasons. Both views place public reasoning at the core of politics in conditions of diversity, and for similar reasons, as we will see.

Although Rawlsian political liberalism is fundamentally normative, it incorporates positive (predictive or behavioral) elements that give rise to the problem of mutual assurance, and create an important link to positive political theory. One is *rationality:* people are assumed to be capable of forming an idea of the good (religious or secular), formulating plans to pursue the good and then acting in a

---

[3] Stephen Macedo, *Why Public Reason?* (under review), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1664085.

[4] Gillian K. Hadfield & Barry R. Weingast, *What is Law? A Coordination Account of the Characteristics of Legal Order.* J. Legal Analysis (forthcoming 2011).

[5] See his magisterial, Robert D. Putnam, Making Democracy Work: Civic Traditions in Modern Italy 12 (1993).

[6] We mean "small c" "constitutional," referring to the design of the basic institutions of a political community, see Walter F. Murphy, Constitutional Democracy: Creating and Maintaining a Just Political Order (2006).

[7] Nicely set out by Paul Weithman, Why Political Liberalism?: On John Rawls's Political Turn (2011).

manner consistent with that plan. Rawls clearly understands rationality as a moral capacity that helps us define and pursue our basic interests, yet it is also a practical constraint on what is achievable in a political community: we must consider it in predicting how people will in fact behave and which institutional arrangements will in fact be stable.

Another positive element is the idea of *stability*. For a political conception of justice to provide the cooperative framework for a political community, it must be sufficiently stable "to persist and gain adherents over time within a just basic structure … if a conception fails to be stable, it is futile to try and realize it."[8] Working out a political conception of justice in conditions of religious and ethical pluralism takes time and work; what citizens need is not a fleeting consensus but a stable, reliable, mutually-assured framework for social cooperation.

Like rationality, stability is thus, a practical constraint on the set of feasible political conceptions of justice. Given the behavioral premise that people are rational and care about and seek to achieve their personal conception of the good (or their "comprehensive doctrine" of truth and value as a whole), we have to anticipate that they will only be willing to participate in and support the institutions and interactions required by a political conception of justice if they can be reasonably assured that they will benefit as a result. To the extent that their participation requires them to bear costs or forego aspects of their own pursuit of the good to benefit others, we expect rational agents to look for assurance that others will do the same for them. As Hadfield and Weingast emphasize, in the absence of a practical, behaviorally realistic, resolution of this mutual assurance problem, a political conception cannot be stable and hence cannot be a reasonable proposal for the basis for political life.

To these positive elements, Rawls adds essential normative ones. In a reasonably well-ordered society, people are not only rational, they are also *reasonable,* that is, willing to seek fair terms for social cooperation given the assurance that others will also do so. Moreover, Rawls hopes that a reasonable political conception can be not merely stable but *stable for the right reasons*: that is, stable because based on a shared public moral conception of justice that stands up to critical scrutiny from all quarters while furnishing a freestanding or independent rationale that "address[es] each citizen's reason, as explained within its own framework."[9] Rawls' theory is squarely normative, and yet the practical concern with stability is never far from view.

This is where public reason and reciprocity come in. The public reasons offered in support of a political conception of justice are a *specific kind of moral*

[8] John Rawls, Political Liberalism 141-42 (1996).
[9] *Id*. at 143.

reasons: addressed to one's fellow citizens with the aim of furnishing a shared justificatory ground for common principles. Public reasons are those that are offered in public as an appropriate basis of fair cooperation, supported by evidence and forms of reasoning that are widely available. The practice of public reason manifests participants' understanding of political relationships as *essentially reciprocal relations* among people of *equal standing*: I will support the exercise of political power over you only according to shared reasons acceptable to you in the expectation that you are similarly motivated to appeal only to shared reasons acceptable to me to justify the exercise of political power over me. The motive to be relied upon for such other-regarding concern, Rawls tells us, is not altruism but reasonableness: the willingness to cooperate with others on fair terms given the assurance that they will likewise do so. Public reason, understood as a reciprocal practice among actual citizens, has both moral and practical (or positive) content: by offering public reasons for laws touching on important rights or questions of justice, we reassure one another of our cooperation on fair terms that we can all share. A reciprocal commitment to fair cooperation justified on the basis of public reasons—a normative commitment conditional on others' normative commitments—thus forms the basis for the legitimate exercise of power according to political liberalism. Public reasons are the building blocks of an autonomous public political morality.[10]

Interestingly, or so we think, the Hadfield-Weingast model predicts that purely rational agents operating under conditions of normative diversity will discern the advantages of cooperating on the basis of general rules informed by a system of shared reasons. Self-interested, but forward-looking citizens, will arrive at a shared logic of cooperation that is independent of each one's personal conception of the good (or idiosyncratic logic). As we will see, the positive model helps explicate the practical role of a democratic practice of public reason: the positive version is not quite *stability for the right (moral) reasons* but rather, as we explain below, *stability for (and via) shared reasons*. The positive model of Hadfield & Weingast usefully draws attention to the inescapable role of ordinary citizens in sustaining just laws, and this informs our interpretation of public reasons' normative role.

B. Decentralized Enforcement: Citizens' Wide Role

Our analysis rests on an idea that is often neglected, namely the idea of decentralized enforcement. In much recent work in political theory and law, the focus is institutional

[10] As Rawls puts it: "Fair terms of cooperation specify an idea of reciprocity, or mutuality: all who do their part as the recognized rules require are to benefit as specified by a public and agreed-upon standard." *See* John Rawls, Justice as Fairness 6 (2001).

and the object of concern is the coercive power of modern states. The exercise of this coercive power over individuals is said to give rise to the problem of political legitimacy: the need to justify this power to the individuals who are subject to it. The popular uprisings against authoritarian rule across much of the Arab world in the spring of 2011 remind us quite dramatically, however, that political regimes that rely exclusively on *coercive force* to enforce the law may prove to be illegitimate and quite fragile. A just regime depends not only on the exercise of power by a centralized authority, but more pervasively, on individual citizens' free and willing "policing" behavior, broadly understood to include very many forms of support and encouragement, or disapproval and censure.

Citizens who criticize and shun, or simply censure, those who violate the rights of others are engaged in policing behavior, as are (unfortunately) citizens who engage in mob violence against those who are exercising their (putative) rights. In either instance, individuals can feel secure in their rights only insofar as they can count on the widespread support of their fellow citizens. This is the crux of our claim for the practical importance of the reciprocal practice of public reason: by giving public reasons and demanding them from others, we hold one another accountable for our subscription to shared standards and we reassure one another of our joint commitment to the shared project of public justice. By taking seriously the practical constraints of rationality and stability, and drawing on the insights of positive theory, as developed by Hadfield and Weingast, we gain a clearer understanding of the central importance to justice of citizens deploying public reasons as ways of manifesting their support for fair cooperation.

We put the role of citizens in supporting justice at center stage: justice, to be effective, must figure in citizens' day-to-day interactions. The security of citizens' rights, especially members of minority groups, depends upon widespread support for those rights. African Americans are not secure in their equal liberties to use public accommodations, women are insecure in their equal standing in the workplace, gay and lesbian students are not equal in the security of their persons, unless they can count on their fellow citizens to join in support of their just entitlements: not only at the polls on election day, but on street corners, in associations and workplaces, and at restaurants and other public accommodations. Our analysis thus connects with those who have argued that citizen play a crucial role in sustaining democratic justice.[11]

---

[11] *See, e.g.,* Stephen Macedo, Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism 1-2 (1990); William A. Galston, Liberal Purposes (1991). More broadly, see the important study of the role of decentralized political activity bolstering democratic capacity in Josiah Ober, Democracy and Knowledge: Innovation and Learning in Classical Athens (2008), an account from which we have learned much.

The emphasis on decentralized enforcement is a central feature and key insight of the Hadfield-Weingast model. They demonstrate that the characteristic features of legal order and the rule of law, securing compliance with rules that promote individual and collective welfare, may be attributable *not* principally to enforcement by a centralized authority but to the need to coordinate and support widespread participation of individuals in decentralized enforcement mechanisms. Such mechanisms include overt collective punishments such as boycotts, protests and refusals to deal with someone with a reputation for rule violations, as well as more subtle and self-mediated mechanisms such as shame or guilt. Decentralized punishments pose two key challenges in a rational actor model. First, participating in the punishment must leave the participant better off, over all, than not participating. Second, in light of diverse interpretations of what conduct is wrongful and hence what conduct warrants punishment, participants need a coordinating device to ensure that they engage in punishments if and only if sufficient others will do the same. Hadfield and Weingast argue that the latter constraint sets up the role for a common logic of wrongfulness to guide individuals in assessing what is right or wrong, punishable or not. The former, incentive, constraint is then met only if this common logic is sufficiently protective of individual participants' interests to improve individuals' well-being relative to the world in which coordinated punishments fail.

The Hadfield-Weingast model does not distinguish between public and private rules or enforcement. The rules it contemplates may be concerned only with mundane transactional details and enforcement may be directed exclusively to securing the benefits of a private deal. So the common logic that secures compliance by coordinating and incentivizing decentralized enforcement efforts in their model is not necessarily concerned with the kind of fundamental political rights or the justification of the exercise of collective state power that are the concern of Rawlsian justice. Thus, we should not be understood to be claiming that all of the decentralized enforcement efforts that are deployed in a society are subject to the demands of public reason; only that a public and shared logic may undergird coordination of such efforts, regardless of the nature of the rule at stake. Rather our goal is to explore how public reason—whether restricted in scope to the justification of rules affecting fundamental rights or extending to exercises of collective power to enforce compliance with more mundane legal rules—can be illuminated by an appreciation of the instrumental role of a public and shared logic that supports the participation of diverse peoples in decentralized enforcement of law.

The key assumption of the positive model that helps us understand the importance of public reason is the recognition that in order to coordinate the participants in a community dependent on decentralized enforcement of rules and principles expressing judgments of right and wrong, the common logic must be

publicly accessible. Indeed, in game theoretic terms, it must be *common knowledge*: everyone must know that everyone knows that everyone is looking to the same logic, and reaching the same (or similar) conclusions, in order to decide what is wrong and what will call forth collective punishment. The requirement that a form of reasoning about right and wrong—what a community considers punishable and what it considers permissible—be common knowledge gives us the central role of *public* reason: a system of reasons that all can participate in and in which participation can be manifest to all.[12]

Along the way we also suggest that the model we advance helps us understand the shortcomings of the "convergence" conceptions of public justification that some argue is superior to the sort of "consensus" model offered here. Convergence theorists, such as Gerald Gaus and others,[13] allow that it is important for citizens to converge on just laws, but they argue that we do not need shared reasons or a shared public political conception. Our analysis helps clarify the practical work done in a diverse political community by the availability of shared forms of reasoning.

This Article is organized as follows. In Section I we lay out the normative conception of public reason in political liberalism and identify its key elements. Section II presents an account of Hadfield and Weingast's positive model of the role of a common logic in achieving equilibrium in a regime dependent on decentralized enforcement by citizens. In Section III we draw out the structural similarities of the positive and normative models, showing how the twin constraints of rationality and stability suggest a key role for public reason as an essential coordinating device in society that depends on decentralized support for the political conception of justice. In Section IV we explore how the positive and normative accounts might interact, arguing that the richer attributes of the normative view concerning *individuals*—not only rational but also reasonable—and *stability*—not merely stable for shared reasons but stable for right reasons—can be understood as playing a positive role, that is, as augmenting the positive theory. The stability of a regime based exclusively on the behavior of conventionally rational individuals may furnish a basis for the evolution of a more robust form of stability. This more robust form of stability is stability based not merely on mutual advantage and an instrumental commitment to a common logic of coordination, but on internalized values of reciprocity and reasonableness that support fair cooperation for its own sake and not merely instrumentally. A society populated by reasonable individuals is one that enjoys widespread voluntary, rather than enforced, fidelity to the political conception of

[12] On the importance of common knowledge, we have benefitted from Michael Suk-Young Chwe, Rational Ritual: Culture, Coordination, and Common Knowledge (2001).

[13] *See infra* note 48.

justice. In a concluding postscript we link the preceding discussion to the politics of rights in modern democracies.

I. Public Reasons as Public Morality

A. Public Reason Decentralized

Public reason addresses the problem of political legitimacy in a diverse society. As Rawls put it:

> Our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to them as reasonable and rational. … And since the exercise of political power itself must be legitimate, the ideal of citizenship imposes a moral, not a legal duty—the *duty of civility—to be able to explain to one another on those fundamental questions how the principles and policies they advocate and vote for can be supported by the political values of public reason*.[14]

Given reasonable pluralism, this view holds, we cannot hope to generate a shared justifying rationale for law based on particular religious or philosophical grounds. We must resort, therefore, to public reasons that are freestanding, and derived from widely affirmed ideas in our public political culture, suitably reformulated into a coherent and defensible conception of justice. According to this view, citizens in a pluralistic liberal society

> are to conduct their public political discussions of constitutional essentials and matters of basic justice within the framework of what each sincerely regards as a reasonable political conception of justice, a conception that expresses political values that others as free and equal also might reasonably be expected reasonably to endorse.[15]

The moral obligation of public reason expresses the reciprocity inherent in the political relations among free and equal citizens in a constitutional democracy: "if we argue that [for example] the religious liberty of some citizens is to be denied, we must give them reasons they can not only understand …. but reasons we might reasonably expect that they as free and equal might reasonably also accept."[16] A

[14] Rawls, *supra* note 8, at 137.
[15] *Id*. at 1.
[16] *Id*. at li.

commitment to appeal to public reasons—reasons that are offered in public as an appropriate basis of fair cooperation, supported by evidence and forms of reasoning that are widely available—expresses an understanding of political relationships as essentially reciprocal.

One aspect of reasonableness, according to political liberalism, is acknowledging the fact that because there is a plurality of reasonable comprehensive religious and philosophical worldviews, it would be unreasonable to base law on any one conception of the truth as a whole. Another aspect of reasonableness is *reciprocity*, understood in normative terms here as the *willingness to cooperate with others on fair terms given the assurance that they will likewise do so*. Reciprocity is a moral motive of a particular sort: neither self-abnegating altruism nor mere self-interest. The reasonable person committed to reciprocal cooperation on fair terms expects political institutions to advance his or her interests fairly along with others. Reciprocity—as a normative commitment conditional on others' normative commitments—thus forms the basis for the legitimate exercise of power.

In developing the argument for why citizens in a pluralistic society should offer public reasons that satisfy the criterion of reciprocity, Rawls appeals to the idea of "stability for the right reasons." For a political conception to provide the framework for a political system that honors the limits imposed by the obligation of public reason, it must be sufficiently stable "to persist and gain adherents over time within a just basic structure."[17] The positive or practical aspect of this requirement is evident: "if a conception fails to be stable, it is futile to try and realize it."[18] A fleeting consensus cannot provide a stable, reliable framework for political interaction. But as the basis for a system of fair cooperation, stability must also be understood in normative terms. Not just any stable solution will do. Stability must be secured by giving all citizens a basis for continued support of the political conception of justice that "address[es] each citizen's reason, as explained within its own framework."[19] The public reasons offered in support of a political conception of justice are a specific kind of moral reasons: addressed to one's fellow citizens as furnishing a shared ground for fair cooperation that each might endorse so long as others also do so.

This account of stability rests on two complex ideas, which together allow us to reconcile the need for common principles and institutions with the "permanent fact of reasonable pluralism." First, is the idea of a *shared political conception of justice* formulated in public reasons. Second, is the idea that the political

[17] *Id.* at 143.
[18] *Id*. at 142.
[19] *Id*. at 143.

conception may gain the support of an *overlapping consensus* of the many religious and philosophical comprehensive doctrines that citizens espouse. We take these two key elements of political liberalism in turn.

First, then, is the idea of a shared political conception of justice formulated in public reasons. We formulate a public conception of justice on the basis of reasons and evidence that we think others ought also to accept. The right sorts of reasons will include the reasonably settled findings of science (among those who examine them carefully), and moral reasons whose force does not depend on embracing a particular religious or philosophical framework. The raw materials for a shared conception of justice are drawn from the public culture of decent constitutional systems: we look for moral convictions that have widespread currency but that are also good candidates for being defensible, sound, and capable of standing up to criticism over time. The wrongness of racial and other forms of discrimination, for example, is one good place to start, as is the importance of religious toleration.[20] We knit such convictions into a coherent and powerful system of ideas—a public political conception of justice. Rawls develops his political conception of justice based on the idea of society as a system of fair cooperation among free and equal citizens understood as having two basic moral powers (for a conception of the good, and a conception of justice). In effect, his two principles of justice unpack and argue for the normative power of a specific conception of those familiar ideas (fair cooperation, citizens as free and equal, etc.). The account is meant to resonate deeply with our political culture but also stand up to critical scrutiny as an account of our basic political interests.

Second, comes the idea of an overlapping consensus. Political liberalism, as already noted, takes it as a fact about the modern world that under conditions of freedom people will come to espouse a wide range of ideas about what is true in matters of faith and philosophy. Pluralism with respect to our deepest and broadest religious and philosophical conceptions is a permanent condition and not unreasonable. An overlapping consensus models the idea that we can accept this pluralism while nevertheless supporting shared principles of justice. Insofar as our differing "comprehensive conceptions" of truth and value allow us to endorse an independent shared political conception of justice, and each of us conceives of our differing rational aims (our highest ideals and widest conceptions of truth and value) as supportive of or congruent with the shared conception of justice, an overlapping consensus obtains.

In substance, Rawls defends the priority of a list of basic rights, a guarantee of what he calls fair equality of opportunity, and a further principle arguing for a fair

[20] *See* John Rawls, A Theory of Justice (1999).

sharing of the benefits of social cooperation (so that the least well off group tends to be maximally advantaged). The resulting theory of ideal justice is controversial: it is not simply a report on what most Americans think, but rather, a critical account that claims to be the best interpretation of core ideas in our public tradition. However, we can also identify a wider "family of liberal political conceptions" containing some shared basic features: an agenda of basic rights whose protection has special priority, and the public provision of at least a basic array of social welfare institutions. It is realistic to expect all major groups in politics to at least espouse one version or another of the family of reasonable political conceptions so defined.[21]

Importantly, the public political conceptions of justice are simply the overlap or common denominator shared by all of the reasonable comprehensive doctrines that exist in society. If our shared account of justice was merely this overlap, then it would be contingent on the happenstance of common elements among existing doctrines or the particular configuration of power among competing views. Rather, the shared political conception is worked out as a *freestanding* conception that is capable of gaining the continual endorsement of individuals and groups who entertain "conflicting and incommensurable"[22] conceptions of the good. If we failed to formulate our political conception as freestanding, it would be dependent upon the current shape of our own religious or philosophical views, or, if formulated as a compromise among the major such conceptions in society, it would be dependent on the current balance of forces among them. In those cases, our commitments to fair terms of cooperation would be contingent and fragile: my commitment would depend on religious or philosophical standards of truth and sources of authority that most of my fellow citizens will regard as opaque, arbitrary, or false.

Similarly, the political conception of justice is a *complete* conception that is capable of addressing all basic questions of justice in its own terms. My commitment to completeness means that if some difficult public question arises I will not abandon the search for a resolution based on shared reasons *even if* my personal non-public views furnish a ready answer. Here again, the point is to remove contingencies that would otherwise render fragile or tentative my commitment to cooperation on fair terms that I can share with my fellows. We should all regard the public political conception (or one among the family of reasonable conceptions) as an independent framework of thought with a logic of its own: as a freestanding and complete conception that any of us can enter into as equals, in spite of our differences of faith and personal ethics, to work out how our shared political morality should be interpreted, applied.[23]

---

[21] Rawls, *supra* note 8, at xlix-l.

[22] *Id*. at 135.

[23] The public political conception is not entirely insulated from the influence of our personal conceptions of truth and value. Each of us must regard the shared project of constructing a mutually

Both positive and normative elements come together to compose these core ideas of an overlapping consensus and a shared freestanding and complete political conception of justice. These include a positive requirement of practical stability: if a political conception of justice cannot, in fact, gain the endorsement of a diverse set of individuals then it cannot plausibly form the basis for a well-ordered society. This requirement of practical stability is built on a further positive consideration, namely the behavioral fact that normal human beings are rational: they form *and pursue* their own conception of the good life. Rawls recognizes "human nature and its natural psychology" as constraints: "We cannot say anything we want [about viable conceptions of persons and their basic interests], since the account has to meet the practical needs of political life and reasoned thought about it."[24] It makes little sense, therefore, to propose a political conception that rational individuals cannot see as consistent with reasonable conceptions of the good *in light of* the shared project of sustaining a well-ordered and just society. A public conception of justice may become a practical possibility only after generations of political development. In the early stages of a given political order agreement is liable to be possible only on a narrower set of extremely urgent shared commitments: say, to peaceful co-existence, and procedures for resolving disputes.

The persons to whom the theory of justice is addressed under political liberalism, are not exclusively rational; they are also *reasonable*. They "desire for its own sake a social world in which they, as free and equal, can cooperate with others on terms all can accept."[25] Likewise, the political conception devised by such reasonable agents should be one that is not merely stable as a practical matter, but rather, *stable for the right—shared public moral—reasons*. The aim is not merely an order based on a particular concatenation of social and political forces with state-enforced penalties to induce compliance,[26] nor an agreement on procedures alone.

The positive concepts of rationality and stability are thus both modified by a normative concept of reasonableness that reflects citizens' desire to cooperate with one another on terms all can publicly recognize as fair. What is reasonable operates

---

acceptable public political morality from the standpoint of our own values and beliefs as a whole: we endorse a version of that shared morality in light of everything we believe in. This naturally leads to some "personalization" of the public view. Realistically, the best we can hope for, as Rawls says, is that most citizens will converge on one among a family of reasonable liberal conceptions, with some core liberal and democratic features. That is sufficient. All of this is consistent with a presumptive or provisional commitment to advancing and abiding by a public political conception of justice that is understood to have an autonomy and integrity of its own, given its public political role. See Rawls's discussion of the "three stage sequence," in Rawls, *supra* note 8, at 385-95.

[24] *Id.* at 87.

[25] *Id*. at 50.

[26] *Id.* at 142.

as a constraint on unalloyed rationality: those who are rational but unreasonable pursue their private interests or comprehensive doctrines, perhaps recognizing the need to bargain with others for the sake of one's own aims, but without sufficient regard for fairness towards those with whom one shares a social world.

Political liberalism's reciprocity, Rawls insists, is "not the idea of mutual advantage."[27] Mutual advantage recognizes only the individualized pursuit of the rational and not fundamentally social, efforts to secure justice among those recognized as free and equal. Stability for the right reasons requires that the political conception be one that is robust to changes in the distribution of power among comprehensive doctrines;[28] it is not based on merely the current balance of forces, or a contingent compromise that is liable to being overturned as the relative power of groups shift. It is, rather, based on a principled commitment to cooperation on fair terms. As such, it should be stable across many future contingencies, and the ebb and flow of power among groups, "its form and content are not affected by the existing balance of power between comprehensive doctrines. Nor do its principles strike a compromise between the more dominant ones."[29]

The constraints of reasonableness and reciprocity are crucial for working out a just basis for a diverse political community. They help define the *public* relations of citizens:

> It is by the reasonable that we enter as equals the public world of others and stand ready to propose, or accept, as the case may be, fair terms of cooperation with them. . . Insofar as we are reasonable, we are ready to work out the framework for the public social world.[30]

A political conception of justice can only be worked out among those who self-consciously recognize the challenges of identifying a shared conception in the face of such diversity of views. It would be unreasonable for a person, who accepts the fact of reasonable pluralism and yet desires to live in a peaceful society in which groups no longer contest for dominance over others, to insist on principles of justice that simply reflect his or her own particular comprehensive conception of what is true.

Because reasonableness and reciprocity are fundamentally public concepts, the obligation of offering *public reasons* is also fundamental. The constraints on the exercise of rationality cannot be satisfied without a shared public logic of what is fair and reasonable. The right reasons that support stability are those that satisfy

---

[27] *Id.* at 17.
[28] *Id.* at xliii.
[29] *Id.* at 141-42.
[30] *Id.* at 53.

the criterion of reciprocity: they express "political values that others as free and equal also might reasonably be expected reasonably to endorse."[31] Public reason is the dynamic medium through which we construct and confirm, collectively, the set of values that satisfy the normative constraints of reciprocity among reasonable citizens. Public reasons do important practical work in a system that depends on decentralized enforcement by furnishing grounds for mutual assurance that makes actual participation in enforcement more likely.

Political liberalism's normative reciprocity as expressed in the obligation of public reason is not to be confused with reciprocity as a purely positive concept, which is rooted in predictions about actual costs and potential benefits given a particular distribution of citizens' preferences. Nevertheless, the positive logic of coordination, based on reciprocal efforts to enforce a set of value-creating rules and principles for behavior, helps us to understand the practical work that public reason can do. We turn to such a positive account now.

II. A Positive Model of Decentralized Enforcement

In political liberalism, the normative and the positive are complementary, and both are necessary. Thus, the concept of reciprocity, Rawls tells us, "lies between the idea of impartiality, which is altruistic (as moved by the general good), and the idea of mutual advantage understood as everyone's being advantaged" as evaluated by each individual's own conception of the good.[32] Similarly the idea of the person employed by political liberalism includes capacities for *both* reasonableness and rationality and neither concept can subsume the other. The persons who sustain the stability of the political conception of justice cannot be exclusively rational: they must be motivated by more than the pursuit of their personal conception of the good: they are prepared to take on the shared project of "work[ing] out the framework for the public social world."[33] Citizens who actively support the political conception of justice have what Rawls' calls a "reasonable moral psychology": they are motivated by the desire to act as the conception requires; "when they believe that institutions or social practices are just, or fair (as these conceptions specify), they are ready and willing to do their part in those arrangements" provided there is reciprocity, that is, "reasonable assurance that others will also do their part."[34]

But we are left with a question. On what grounds should we expect that people with a reasonable moral psychology, and the necessary conception-dependent

[31] *Id.* at l-li.
[32] *Id.* at 50.
[33] *Id.* at 53.
[34] *Id.* at 86.

desires, will predominate in a liberal political community? Rawls' answer to this question appeals to the educational effects of well-ordered institutions: "those who grow up under just basic institutions acquire a sense of justice and a reasoned allegiance to those institutions sufficient to render them stable."[35] This makes sense if we assume that we already have just institutions that secure general support. But it begs the question of how reasonable people and just institutions arise. For a population of people cannot grow up under basic just institutions, unless those institutions are already reasonably just and stable. Since justice depends on citizens' willingness to participate in reciprocal public reason giving and demanding, the question is critical.

Rawls suggests a positive account of how the stable, just world in which citizens are constituted as reasonable citizens by their education and participation in just institutions might come about. He talks at one point about the transition from a *modus vivendi* (essentially, a peace treaty based on a balance of forces) to a "constitutional consensus" and then ultimately to a more principled public moral conception, a well-ordered society.[36] Here the main focus is on how people with differing and conflicting comprehensive doctrines might at first establish terms of co-existence based on a balance of forces, and later on a richer set of procedural values (rights of participation and representation) that help stabilize cooperative relations. The standards internal to a procedural (or constitutional) consensus may attract greater support and allegiance over time and develop a life and logic of their own. A shared political morality of principles and conceptions, may eventually come about, one that treats reasonable persons with differing conceptions of the good as equally entitled to equal liberty and distributive fairness, supported by an overlapping consensus: a shared commitment to a public morality of justice no longer dependent on the current distribution of power.[37]

As a positive account, however, this is suggestive only. It does not explain why a society based on a balance of power will come to shift from an organizing principle of power or interests to one of justice. And indeed, this is a fundamental problem in practice, as many if not most societies have failed to make such a transition.[38]

In this section we build on a model of the characteristics of distinctively legal order developed in Hadfield & Weingast to provide a positive account of the role of

[35] *Id.* at 142.

[36] *Id.* at 158-68.

[37] *Id.* at xxiii-xxxii.

[38] Douglass C. North, John Joseph Wallis & Barry R. Weingast, Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History (2009).

public reason in the transition to a normatively-constituted legal order.[39] Hadfield & Weingast begin with the idea that a legal order is, minimally, characterized by an equilibrium in which behavior is patterned on the basis of a deliberately chosen normative classification of conduct, identifying which behaviors are considered wrongful and which are not. Rawls's well-ordered society is an example of a legal order. In it, reasonable citizens follow a set of rules and principles that establish the basis for fair cooperation between free and equal citizens. The rules and principles provide a normative classification of conduct and this normative classification is understood to be the object of deliberation and choice by recognized processes such as a constitutional convention, judicial decisions and legislation.

Positive political analysis assumes that individuals are rational in the same sense that economists understand it: they are motivated solely by reference to their personal assessment of what is good. They are not motivated by other-regarding concerns to assist fellow citizens when threatened with unfair treatment. But as rational agents they are capable of seeing that the world in which there is coordination to effectively punish certain behaviors and not others might be better for them than one that lacks such coordination. A rational agent, even one who expects others to coordinate, will care about two things in deciding whether to participate in collective punishments coordinated by a common logic. One is the extent to which the common logic converges with the agent's own personal logic: will a system of community punishment coordinated by the common logic help deter enough of the wrongs, and promote enough of the goods, that the individual agent would like to see deterred or promoted? Another is, will participation impose too many costs or require the individual to compromise choices too much? Rational agents must, in other words, weigh the benefits and costs of coordination against what they can achieve in the absence of coordination.[40]

The Hadfield-Weingast model derives an equilibrium in which rational individuals adopt as a set of behavior-guiding principles a common logic that does not perfectly coincide with any individual's personal assessment of the good, but which must, to be stable, reflect to a significant extent how diverse others see the world. Most importantly, the model explains in positive terms how rational individuals will seek to adopt a system of shared reasons and principles that is—like liberal public reason—freestanding, complete, and embraced as a matter of common knowledge. By exploring the connections between this positive model and liberalism's normative account, we attempt systematically to connect positive rationality to a morally richer model of fair cooperation.

---

[39] *Supra* note 4.

[40] Rawls suggests something similar in his stylized "three stage sequence." Rawls, *supra* note 8, at 385-95.

The positive model focuses on an idea often neglected in theories of legal order: the central role played by *decentralized*—as opposed to state—enforcement of the rules that constitute legal order. We refer here to the wide variety of ways in which individuals impose costs large and small on themselves and others who violate norms, or confer benefits on those who comply with norms. Someone who violates a norm might worry about the reactions of others, if detected, or might even feel guilty—and so avoid violations even when the chances of detection are slim. Others who observe violations might engage in behaviors that make the violator feel ashamed. Or they might withdraw cooperation and support—refusing to take in an outlaw or boycotting a local business, for example. Violators may be punished by people passing on negative appraisals of their conduct; thus acquiring a bad reputation, they may have difficulty finding people with whom to trade or socialize. Although centralized coercive enforcement of rules by the state often plays a key role as a backstop for decentralized enforcement, legal systems cannot afford to rely exclusively on the threat of coercive enforcement to achieve high levels of compliance. States that cannot rely on their citizens to cooperate in enforcement efforts must spend substantial resources monitoring and adjudicating behavior and/or employ substantial, disproportionate, penalties to compensate for the likelihood that much non-compliance goes undetected or unproved. Coercion-based legal orders are inefficient and fragile, in positive terms, and morally illegitimate.

Decentralized enforcement poses a fundamental problem of coordination. Most collective punishments require a critical mass of participants to make the costs to an individual manageable and/or to make the threat of punishment effective in deterring non-compliance. Boycotts, protests, and retaliation are effective only when significant numbers participate. Passing on negative information about a trading partner has a greater effect on the partner's reputation—and thus greater capacity to deter bad behavior—if those receiving the information evaluate it in a similar light. Indeed, "gossip" can often be understood as a way of expressing and reinforcing shared norms of conduct, by using the relevant standards to report on or inquire into the conduct of a third party. It typically assumes that one's interlocutors share one's evaluative standards because, if they do not, one's praise or criticism of another may suggest that one's own standards are deviant or improper and this may redound to one's own detriment. Even if the "punishment" mechanism is based on internalized standards, and so voluntary, the evaluation of what a given set of norms require and the maintenance of commitment to those standards depends for many people on how they expect others to respond.

Collective punishment requires individuals to share common beliefs about which behavior warrants punishment. Both potential wrongdoers and citizens, who might participate in punishment, need to make reasonably reliable predictions about whether and when sufficient others will join in punishment. But if, as it seems

reasonable to presume, individuals are diverse in their personal assessments of what is ”wrong”—if they each possess what Hadfield & Weingast call an *idiosyncratic logic* for assessing wrongfulness which is significantly inaccessible to others—how can they coordinate?

There is a second challenge that a system that relies on collective punishment must meet. An agent who knows *how* to coordinate punishment with others—when to criticize someone in trading or other social relations, when to boycott, when to show up at a protest, etc.—still must decide *whether* participating in punishment is worth it. Punishing someone, even oneself, is costly. It may entail giving up a relationship or trading a partnership that is otherwise beneficial, spending time at a protest, or running the risk of harming one's own reputation if criticism of another is perceived as unfair, unreasonable, or unwarranted. Given the preferences conventionally posited in economic models—that is, in the absence of personal preferences that directly value compliance with norms and participation in punishment of others in order to uphold those norms—a rational agent needs to identify a benefit to outweigh the costs of compliance and punishment.

Any stable legal order relying on decentralized enforcement among purely rational individuals with conventional preferences has to solve these two problems: coordinating punishment efforts and providing individuals with a motivating incentive to incur the cost of participating in costly punishment. In this positive approach, stability is modeled as equilibrium. In equilibrium no individual has an incentive to depart from equilibrium behavior—refraining from wrongful conduct him or herself and punishing those who do act wrongfully against others—on the assumption that all others are behaving in the same way.

Ideally, rational individuals would like to see the common logic exactly replicate his or her personal idiosyncratic logic. The most immediately attractive option for a rational individual would be for politics to take its bearings from his or her own comprehensive/ideal viewpoint. But the positive model, like the normative model of public reason, assumes diversity in the form of ethical and religious pluralism. So everyone knows that circumstances do not permit anyone to act effectively as dictator. As a consequence, if—as we presume—a well-ordered and stable pluralistic society depends at least to some extent on the willingness of ordinary citizens to participate in upholding the rules and principles that make order possible, then rational agents know that there is no feasible alternative to seeking the free and willing cooperation of others. The problem is to find a solution that will sufficiently engage the interests of all (or sufficiently many to make punishments effective) to induce them to participate.

In Hadfield and Weingast's model, enforcement can be coordinated among a diverse set of individuals by a common logic. A common logic is a publicly accessible system of rules, principles, and reasons for classifying behavior as wrongful or not.

The function of the common logic is to produce common knowledge classifications of behavior as warranting punishment or not. Looking to the common logic, and knowing that all others are also looking to the common logic, to decide whether to punish particular actions of fellow citizens—to speak badly of them, to withdraw fellowship or trade, to ostracize them or take affirmative steps to protest—individuals are able to predict the participation of others in punishment efforts.

Hadfield and Weingast demonstrate that an equilibrium legal order that solves the coordination and incentive problems of decentralized enforcement can be achieved by an institution that supplies a common logic possessing many of the attributes that we intuitively associate with the rule of law:

- *Public accessibility* and *impersonal reasoning* that allows any agent to implement the logic to reach a common classification.
- *Public reasoning* that allows all agents to observe how the institution implements the logic in new circumstances.
- *Open process* that allows heterogeneous agents to introduce idiosyncratic information and reasoning into public reasoning.
- *Immanent* and *generalizable* principles that allow classification of circumstances that the institution initially cannot anticipate.
- *Unique classifications* that can coordinate expectations in the manner of a focal point. This requires both a single logic and, ultimately, a single classification of particular circumstances.
- Given the potential for ambiguity, an *authoritative steward* of the logic, able to definitively resolve ambiguities and conflicts arising from the implementation of the logic.
- *Generality* that ensures that the logic applies to the circumstances and values of all agents required to punish if punishment is to be effective.
- *Stability* that allows an agent contemplating participation in punishment of a wrong done to someone else to anticipate that the logic will remain the coordinating logic in the future in the event that this agent is the potential victim of wrongful conduct.

It is important to understand how these normatively attractive attributes solve, in a positive sense, the coordination and incentive problems in a system with decentralized enforcement. Consider the coordination problem first. Clearly, for a common logic to generate common knowledge classifications—providing a basis for individuals to predict when others will participate in punishment—it needs to be *publicly accessible*. But this requires more than mere publication of the relevant

standards. An *impersonal* system of reasoning that is invariant to the identity of the person engaged in the reasoning furthers the aims of public accessibility and hence common knowledge by ensuring that any of the participants with diverse conceptions of the good can take up the common logic. In addition, *uniqueness*—which can be generated by an *authoritative steward* such as a supreme court—helps people to coordinate on assessments of conduct by reducing ambiguity about what the common logic classifies as wrongful. Of course, the steward, to be successful, must offer interpretations of the common logic that participants are prepared to go along with. Authoritative stewardship is not a substitute for decentralized enforcement but a way to facilitate coordination on shared standards of conduct.

Now consider the incentive problem facing potential punishers in a decentralized system. Solving this problem requires that individuals deciding whether to participate in punishing wrongs defined by the common logic find in it assurance that their own circumstances and concerns would be considered if they were personally involved. If they are going to incur the cost of helping to deter wrongs against a fellow citizen, as rational agents, they will be looking for assurance that wrongs against them will also be addressed in turn. A classification system that is *open* to individual input, in the sense that it will admit consideration of an individual's perspective in an effort to reconcile personal concerns with the common logic, serves this purpose. This is analogous to individualized due process, and the guarantee that individuals or groups who disagree with the way their rights are being interpreted will have their day in court. Similarly, it is also often helpful if the common logic is not altogether rigidly formulated in terms of specifics: insofar as it is, at certain points at least, formulated in terms of abstract and *generalizable* principles it may be better able to be adapted to new and unforeseen factors and considerations and thus to accommodate the interests of a wider group of individuals.[41] A system of classification that is general—meaning that it is addressed to the concerns of all of those who are essential to an effective decentralized enforcement system—also serves to give individuals confidence that their interests will be taken into account in devising the protections and freedoms of

[41] Of course, there is a well-known discussion of the relative advantages of specific and rigid, versus abstract and general, constitutional provisions. On the one hand, if small states in a federal system, for example, require reassurance that they will have adequate representation, it may not be enough to guarantee them "adequate" representation in a constitutional document. They may, rather, want it to be specified that every state gets two senators regardless of population. On the other hand, with respect to other areas of controversy it makes more sense to rely on abstract standards that allow for contestation and adaptation over time, concerning, for example, what constitutes an "unreasonable" search or seizure. For an excellent discussion, see Christopher L. Eisgruber, Constitutional Self-Government (2001).

the common logic. *Stability* helps to generate an expectation that the rules are not going to change significantly between the time an individual makes decisions about helping to punish others and when that individual may need assistance himself or herself; this supports the individual's incentive to participate in punishment.

Rational agents know that their own personal ideals (or idiosyncratic logics) are not candidates for being a social standard for coordination, due to the fact of diversity. When they decide to participate in enforcement of the rules established by the common logic, they are doing so in the spirit of rational accommodation: the common logic must protect their interests sufficiently but it cannot be expected to reflect their personal preferences in every case. Rational agents in this model are willing to subscribe to a system of reasoning that must be, if it is to do what they want it to do, *independent* of their own personal ideals. Like public reason on the liberal model, the common logic must have logic of its own.

Equilibrium in the Hadfield-Weingast model depends on ongoing communication among agents to confirm that the common logic continues to be sufficiently attractive to each to garner their support. In formal game theory terms, this is what makes the equilibrium "sub-game perfect"; informally, it is what makes the collective threat to punish credible. If an agent decides not to punish action that is classified as wrongful by the common logic, then other agents interpret this to mean that the agent no longer supports the common logic. They in turn will be less likely to punish because they do not wish to find themselves punishing when there are too few others prepared to join in. Individuals in the positive model must communicate publicly to one other their ongoing commitment to the common logic.

Thus, according to the positive model that we draw on here, purely rational agents with conventional preferences have an incentive to subscribe to a shared system of public normative reasoning so long as they have confidence that it is simultaneously responsive to their concerns *and* to the concerns of diverse others (on whom they depend for a stable system of enforcement that generates legal order). The positive model has striking similarities to liberal public reason.

III. Structural Similarities

The agents that populate the positive model are assumed to be exclusively rational. The moral agents that Rawls contemplates have a richer set of motivations that include a desire to cooperate on fair terms given the assurance that others will likewise do so. Notwithstanding these important differences in motivational assumptions, however, there are striking parallels between the positive and normative models. Indeed, beginning with the minimal positive assumptions of rationality and practical stability, the positive model winds up with an account

of institutions and social arrangements that seem conducive to the emergence of reasonable individuals, understood in the moral sense of those willing to cooperate on fair terms with diverse others. Or so we seek to suggest in what follows.

Let us survey the key components of liberal public reason and their correlates in the positive model described above.

A. Public Political Conception of Justice

The centerpiece of Rawls's theory is a public political conception of justice. It is political in the sense that it is worked out in the awareness of its function to secure a stable framework in which reasonable individuals can enjoy fair terms of social cooperation in the conditions of reasonable pluralism.

The common logic in Hadfield and Weingast's model is also worked out in the awareness of its function to coordinate a common knowledge understanding of what is right and what is wrong. It provides the framework for the enforcement of the rules of social life that allow individuals to achieve a better outcome for themselves by coordinating with diverse others.

Our claim is not that the normative account of the public political conception of justice is completely co-extensive with the common logic suggested by Hadfield and Weingast. Important differences exist. Rawls contemplates a political conception of justice that is self-consciously informed by moral reasoning about what is fair for all citizens, in light of an endorsement of the value of accommodating diverse conceptions of the good. The common logic in the positive model is not developed through moral reasoning about fairness: it will reflect considerations of fairness only to the extent that these necessary to encourage others' participation in the enforcement activity that generates benefits for all. In the common logic supported by rationality alone, some may indeed enjoy greater benefits than others and this may be unfair in Rawlsian terms. The key observation is not that the positive model can generate fairness but that it can, surprisingly, generate a common logic that is animated by attention to terms that are acceptable to others even when agents are exclusively concerned with their own welfare, and this is a step in the direction of fairness.

It is also important that both models share a similar understanding of the "public" political conception of justice or logic of coordination. On the normative model, it means not merely that everyone can learn the content of the political conception; the conception is public in the deeper sense of being accepted as reasonable by members of the community as a basis of social cooperation in light of their comprehensive conceptions. The common logic in the H-W model is also public in a dual sense. It must be publicly accessible if it is to function as a coordinating device. But more than this, it must be public in the sense that it is sufficiently convergent with the idiosyncratic logics of those whose participation

is required for stability; it must reflect enough of the interests of individual citizens to garner their active support. The two models contemplates different processes for settling on the public political conception of justice—the normative model contemplates a deliberative process; Hadfield and Weingast contemplate individual parallel decision-making—but both aim to arrive at principles that are publicly accessible and publicly acceptable.

B. Overlapping Consensus

On both the positive and the normative account, a stable common logic (or public political conception of justice) must secure the support of an "overlapping consensus" of the major differing idiosyncratic logics, personal ideals, or "comprehensive doctrines" that citizens espouse. That is, the shared logic of coordination (or fair cooperation) cannot be too far from or greatly at odds with citizens understanding of their rational good (or their "comprehensive doctrines"). In both accounts, citizens must be able to see some compatibility or congruence between the shared public logic of coordination and their personal aims and values.

Individual agents recognize that all face the same incentives: to participate in collective punishment but only if the coordinated result makes them better off according to their personal calculus, or, on the moralized version, is acceptable from people's differing comprehensive standpoints. The *only feasible common logics* are those that enlist the participation of diverse others in a common logic of shared laws that they can regard as reasonable for all. They will seek out a basis for collective action (coordination and collective enforcement) as a "framework for the public social world"[42] that can be endorsed by others in light of their own personal conception of the good. An interest-based explanation provides a positive foundation for stability of a legal order that must exhibit an overlapping consensus analogous to the one that Rawls posits if it is to enlist the willing support of a sufficient number of citizens.

Again, it is important to emphasize the limitations of the positive account here. Participants in the positive model must attend to others' interests (including their convictions about what is fair) to secure their participation in the enforcement of rules that benefit the community. But the impetus toward inclusion is limited on the positive account: those who are not needed nor have no capacity to punish others for rule violations will have no impact on the content or stability of the common logic, and so their interests may be ignored and they may be treated unfairly. The constraints on the common logic—generality, openness, and accessibility, for example—are constraints only with respect to those who are needed for enforcement.

---

[42] Rawls, *supra* note 8, at 53.

The positive model does not get us all the way to an overlapping consensus of all reasonable comprehensive doctrines that citizens espouse. But it does provide a foothold for understanding how purely rational agents can come to organize their social lives on the basis of a common conception of right and wrong supported by an overlapping consensus among those who differ on other matters.

C. A Freestanding and Complete Conception of Justice

In order to secure an overlapping consensus, Rawls emphasizes that the public political conception of justice must be freestanding and complete if it is to be stable. It cannot be derivative of or dependent on any particular comprehensive doctrine. For Rawls the reason is normative: the political conception of justice must, if it is to be stable for the *right* reasons, not reflect a particular balance of power among existing comprehensive doctrines. Nor can it be dependent on the particular configuration of beliefs in existing doctrines. If it is to meet the normative criteria imposed by respect for all citizens as free and equal, it must treat all doctrines equally. Thus, it cannot be controlled or filled in by any single doctrine.

The common logic in Hadfield & Weingast is also freestanding and complete, although not with respect to *all* citizens, but with respect to those whose participation in enforcement is required to achieve stability. Rational agents will be unwilling to incur the costs of supporting a common logic and participating in enforcement if the logic is subject to the control of any particular set of individuals. This is why Hadfield and Weingast argue that equilibrium legal order can be supported by an institution that provides a common logic that is general, impersonal and open to consideration of how idiosyncratic reasoning can be reconciled with common reasoning. Rational agents will be more likely to persist in supporting a common logic if its principles are generalizable and so capable of responding to their idiosyncratic and potentially unknown or evolving interests. This common logic must be regarded as complete so that its gaps are not subject to being filled in by ad hoc reasoning based on purely personal considerations.

The positive model suggests an account based in equilibrium for the stability of a shared logic that is freestanding and complete. In a world in which people have a choice about whether to support a common logic or not, there is potential competition among alternative logics. A common logic that better achieves the characteristic of standing free from the control of any one set of individuals and their idiosyncratic views is more likely to be robust to the vagaries of change and to secure the type of support necessary to persist.

D. Reasonable Moral Psychology

One of the challenges for Rawls's theory is an account of the motivation of citizens to support just institutions. Rawls posits the development of a reasonable

moral psychology, in which individuals' rational self-interest is regulated by an internalized concern for treating others fairly. As for how the preferences of the rational individual come to be regulated in this way, he suggests only that individuals who are raised in just institutions will come to value fairness and justice for its own sake. As we have already noted, this raises the puzzle of how the just institutions that can produce moral agents come about in the first place.

The positive model suggests a relationship between rationality and a reasonable moral psychology. The model presumes that individuals are exclusively self-interested in the conventional sense: they do not place value per se on respecting the equal worth and dignity of others or supporting just institutions for the sake of justice alone. But these agents nonetheless come to *act as if* they valued fairness for its own sake: willingly punishing wrongs done to others, wrongs judged by a public conception of what is right or fair. With an eye to the future benefits of cooperation, they can see that they are personally better off when social life is organized on the basis of an effective shared logic. These are not mere *quid pro quo* exchanges, each of which is expected to be mutually advantageous. Participants recognize that the common logic will in some instances depart from their private evaluation of what is best. In this sense, rational agents are willing to act reasonably: not because their preferences have been modified, but because they recognize that the value of enduring cooperation on shared terms, and also the fact that diversity precludes cooperation on the terms that they might personally ideally prefer.[43]

Again, we emphasize the limits to this structural similarity. The rational agents in the positive model act as if they valued treating others fairly in the sense of upholding rules that protect the interests of at least some others, but only those others who play a role in effective punishment and stable cooperation. There is no push toward greater inclusion based on a moral recognition of other persons as equals. Moreover, even within the circumscribed circle of those who count in the positive model, the content of the shared logic is limited to what is necessary for a stable equilibrium and this will fall short of the demands of justice.

Yet, it is important to recognize that the positive model goes beyond the mere striking of a bargain between agents with more and less power. The positive common logic must secure expectations about how the shared scheme will play out over time, in new and unexpected circumstances and in light of interests and concerns that particular individuals may now find unfathomable. It is not merely a static bargain, a divvying up of the spoils of cooperation. It is a system for generalized reasoning about shared interests. This seems a plausible and necessary

[43] *See infra* Section IIIE (examining this issue further).

first step to the development of the more robust and extensive concern for others that Rawls envisions.

E. Stability for Shared Reasons but not Right Reasons

We might think of the positive model as displaying *stability for shared reasons*: the reasons arise from the practical/prudential need for a shared logic to stabilize coordination for the sake of mutual advantage. "Shared reasons" on the positive account are not Rawls's "right reasons." Fairness as such and other moral notions do not enter the purely positive analysis: rational individuals are not as such motivated to act based on Scanlon's reasons that others could not reasonably reject.[44] Although rational agents, as we have seen, will have an incentive to participate in a framework in which they act to protect the interests of (some) others and subscribe to a system of public moral reasoning that treats people (at least somewhat) fairly, the stability of the equilibrium is never fully removed from individual calculus. Rational agents in the positive model do not value fair cooperation as such, unlike Rawls' reasonable agents. If private valuation shifts so that the equilibrium secured by the common logic no longer generates private benefits to justify the cost of supporting the institution, then individuals will withdraw their support. This threat of withdrawal, in fact, is a key reason why Hadfield and Weingast suggest that the institution providing the common logic will have an incentive to ensure that the logic is as open and compatible with as many idiosyncratic views as possible.

While there are structural similarities between the two models, we do *not say* that the positive model is the basis for a genuine public morality. The common logic must be general and open with respect to the interests not necessarily of *all* citizens, but only those that are essential to the efficacy of decentralized enforcement. If women and slaves, for example, lack capacity to deliver penalties—if they have no discretion to withhold benefits and so no threat to punish by withdrawing cooperation, say—then their interests can be excluded from the common logic. Indeed, if some people who are important to the efficacy of decentralized enforcement have a preference for dominating marginalized groups, or if there are payoffs to be garnered as a consequence of their exclusion, rational agents may have a positive incentive to exclude and oppress others. The equilibrium could still be stable, but here it is plain that this is not stability for morally defensible reasons. The specific equilibrium in positive theory depends on the configuration of power and preferences in society. But by attending to the problem of coordinating and incentivizing participation in decentralized enforcement, the equilibrium in the positive model nonetheless goes beyond a static bargain between the powerful; it contemplates the development of a

[44] *See* Thomas M. Scanlon, What We Owe to Each Other (1998).

shared means of abstract reasoning about interests in common and this seems a step toward genuinely other-regarding moral concern.

F. Reciprocity as a Virtue of Citizenship

The positive model provides an account whereby *rational* agents who seek to coordinate collective punishments in order to secure their own well being will, as much as possible, practice a form of reciprocity.[45] This is not, as already mentioned, a narrow form of *quid pro quo* reciprocity whereby individuals make implicit bilateral contracts in which each participant expects a return after each exchange. Rather, this is a generalized reciprocity, whereby people aid their fellows in the expectation of more diffuse compensation as the beneficiaries of similar cooperativeness from others in the future as needed.[46] This robust form of generalized (or strong) reciprocity only becomes possible based on the rich shared logic and institutional framework that expressly generates normative principles for when others are entitled to the cooperation of their fellows, securing as much of their personal interest as is compatible with similar assurances for all. Although self-interest limits the extent to which rational individuals are willing to contribute, Hadfield and Weingast suppose that participants are willing to perform now in the hope of future returns, so long as there is general support for a cooperative logic. This system requires public affirmation of the common logic and appropriate actions. A commitment to joint enforcement of the terms of cooperation takes place in the shadow of the future, rooted in a common logic that must be impersonal and freestanding if it is to endure. Rational actors are thus led to consider (at least some) others' perspectives as well as their own, and to affirm this publicly.

The reciprocity of positive theory is not moral: but to some degree participants act *as if* they are so motivated, at least vis-à -vis those whose participation is needed to sustain the shared logic of cooperation. Participants in the positive model exhibit a kind of purely rational reasonableness, or a prudential reciprocity: they recognize that it is advantageous for them to sustain a shared logic of cooperation that sufficient numbers of others will also be prepared to support, and so they compromise with others. They treat others' interests as deserving of their support, in the expectation that others will do likewise. They treat the shared logic as a normative system—one generating reasons for compliance and standards of assessment—though not a moral system (more on this distinction below).

---

[45] See the helpful discussion in Brian Barry, Justice as Impartiality chs. 1-3 (1996).

[46] Bowles and Gintis call this "strong reciprocity." Samuel Bowles & Herbert Gintis, *The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations,* 65 Theoretical Pop. Bio. 17 (2004).

G. Public Reason and the Virtues of Listening

The positive model helps to make clear the important practical work done by an enduring system of public reason. Indeed, the persistence of the equilibrium in the Hadfield-Weingast model depends on participants' willingness to engage in both punishment of wrongdoers and appropriate *communicative* action. The social order is characterized by dynamic pluralism, and this helps generate the need for a shared logic of cooperation. Conditions change, as do people's understandings of their own values and those of others. By participating in collective punishment, consistently over time whenever the need arises, and doing so on the basis of an avowed commitment to a common logic (or public conception), individuals signal to their fellows that the common logic continues to be acceptable to them. Failure to participate degrades the stability of the equilibrium, causing others to suspect that coordination is fragile. This encourages wrongdoers and discourages enforcers, who—even if they themselves continue to find the common logic acceptable—are unwilling to bear the costs of punishment with a dwindling number of compatriots. The common logic fails in its coordinating function if, by word and deed, trust in others' willing participation breaks down.

While individuals' idiosyncratic reasons (or personal conception of the good) cannot serve a coordinating function, this does not mean that they have no place in public discourse. Public reason on the positive model does not entail a negative obligation to disavow or remain silent about idiosyncratic views; rather, it entails only an affirmative obligation to take on the task of integrating those views with a common language of reasons.[47] And this is why idiosyncratic reasons are insufficient to support the stability of the well-ordered politically liberal community. Even if I share the idiosyncratic view that you expound, my interest in continued support for our shared framework rests on my confidence that a justification can be offered that is sufficiently acceptable to *all* those who play a role in supporting the framework. I need to hear that you are on board for our societally shared logic of coordination, and I will likewise do my best to demonstrate my own commitment. This is what the freestanding and impersonal common logic makes this possible. We can predict that rational agents will appeal to it in order to ensure its continued vitality as the basis for valuable social cooperation.

H. The Practical Advantages of Publicly Reasoned Consensus

Some critics—the "convergence" theorists mentioned earlier—have noticed that the problem of legitimacy might be addressed adequately without public reason. Scholars such as Christopher Eberle and Gerald Gaus reject what they call "consensus" conceptions of legitimacy—depending on a *shared* public rationale

[47] We are indebted to Nir Eyal for this point.

for principles of justice or a constitution—in favor of "convergence" conceptions that allow for *plural rationales* to undergird a system of law.[48] These critics or revisionists agree that a system of laws should be justified to everyone, but insist that it need not be the same justification for everyone. So long as different religious and ethical communities have their own justifications that converge in support of a system of law there is no need for a *shared basis* or a *common logic* of *public reasons*.

On the face of it, the critics have a point: it is possible in principle that the legitimacy of law could be established on the basis of a plurality of differing rationales. So the question is what, if any, role public reason plays? We think that the practical work of public reason is better understood if we attend to the key practical constraints of rationality and stability: a political conception of justice is unlikely to be stable if it lacks the support of a fairly robust democratic practice of public reason. Given the importance of decentralized enforcement, and the fact that legal rules need to be interpreted and applied in new circumstances in ways that are intelligible and defensible to our fellow citizens, individuals need to know not only how the public scheme is squared with their private views—evaluating how it serves their goals and interests—but also to predict how others will evaluate particular actions as either calling for enforcement or not, and how. The knowledge furnished by mere convergence—which might include the ability to predict that others are stably committed to our current schema of law – does not furnish an adequate framework for extending a complex system of rules and principles into the future.

IV. From Positive to Moral and Back Again: Rational and Fair Cooperation

So far we have seen how various requirements, rooted in positive theory, are structurally similar to parallel ideas rooted in a political morality. We have also seen the gaps between the positive and normative accounts of a shared system of public reasoning. Agents' motivations and reasons for action differ categorically on the two versions. In the positive model, rational agents pursue their self-interest in what Tocqueville called an "enlightened" way: by recognizing the advantages of a stable logic of coordination with others.[49] On the normative account, morally

[48] Christopher J. Eberle, Religious Conviction in Liberal Politics (2002); Gerald Gaus, The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World (2011). For a development of the distinction between "convergence" and "consensus" views, see Fred D'Agostino, Free Public Reason: Making It Up as We Go (1996).

[49] *See* Alexis deTocqueville, Democracy in America (J.P. Mayer ed., 1969).

motivated agents recognize the virtues of setting aside certain long-disputed religious and philosophical issues in favor of principles of fair cooperation that can be justified to, and secure the support of, reasonable fellow citizens with similar motivations. In the positive model, a fair system of cooperation is circumscribed in both scope and content: it takes into account the personal interests of only those who are important to the ongoing stability of the scheme, and it may only be as fair as is necessary to secure the participation of such individuals—some may benefit much more than others. Normative accounts of fairness obviously require more.

On the one hand, the considerable structural similarities between the positive and moral versions suggest ways of thinking about how normatively motivated citizens who are moved by moral considerations as such might emerge from an environment that is governed exclusively by the positive forces of rationality and practical stability. And on the other hand, there might be a "return trip," so to speak, whereby agents who bring a moral sensibility to bear on previously interest-based coordination improve on the outcomes available to rational agents, thus expanding the reach and depth of the common logic that supports fair cooperation and achieves not merely stability but stability for the right reasons

On both the positive and the normative models we are exploring here, law is understood as a human creation that must respond to the claims of diverse individuals by furnishing them with shared reasons for compliance and participation in joint enforcement. On both views, the working of a system of authoritative law depends upon widespread endorsement of the value of cooperation on the basis of reciprocally shared reasons. Both views look toward widespread participation in a shared discursive and justificatory enterprise, as crucial for mutual assurance. Good reasons on both views are "we" reasons: reasons that are good for us as co-participants in a joint enterprise. Both models, finally, regard the systems of shared reasons that each generates as freestanding and complete—autonomous and integral systems of ideas—that are also open to input from citizens' differing idiosyncratic non-public systems of belief. And yet, on the positive model, public justification is conducted in the framework of mutual advantage, while on the normative model, mutual justification is on the basis of public moral claims and a desire for a fair sharing of the gains of social cooperation.

In this section we consider the relationship and interactions of the two logics: How far can a positive game theoretic logic take us in the direction of public morality? And, after we cross the bridge to moral judgment, how might the return trip reveal ways in which the introduction of a moral dimension strengthens—or weakens—the positive logic of coordination? Is progression in the direction of moralized social norms itself suggested by the positive logic?

We raise here the prospect that a common logic of cooperation grounded in mutual advantage could at first become stable for interest-based reasons in a

world populated exclusively by rational agents with conventional preferences. Subsequently, however, agents who participate in a freestanding common logic of coordination may then come over time to make additional justificatory demands on one another. Some might introduce avowedly moral ideas of fairness, equity, and the freedom and equality of participants, into the bargaining, and such categories might gain a grip on participants via the working of what Rawls calls a "reasonable moral psychology."[50] That is, such categories might be recognized as furnishing grounds for claims that require response in the appropriate moral register: based on what fair or equitable cooperation requires, and what is rightfully mine or yours. We ask here specifically about the relationship between positive analysis and liberal democratic political morality.

A. Law, Normativity, and Morality

We have so far elided the distinction between "normative" and "moral" in relation to law; however, we now need to say something about their relation. We acknowledge that these matters are highly contested, the boundaries often fuzzy, and our account is brief.

The positive model predicts the emergence and stability of a normative system (the common logic of cooperation) but not necessarily one that is moral. Normative systems generate standards of conduct that are available in social life as general guides to conduct. Normative systems are necessarily social and evaluative, but they are not necessarily moral; moral systems, such as Rawls's, are a subset of normative systems.

Normative systems may be formal, with institutionalized authorities and explicit rules, such as is the case with law, but they may be informal, like etiquette.[51] Etiquette is fairly though not completely informal, and it is generally not regarded as reflecting moral categories and judgments (or at least serious ones). It involves such matters as not picking your teeth in public, chewing with your mouth open, or reaching across others for food at the table. The reason for "hedging" on the issue of whether etiquette is moral in nature is that its rules often do involve a very general moral aim: showing adequate consideration for others, admittedly in ways that are culturally specific. Other normative systems may be even further from morality. Consider a code of healthy eating—such as the food pyramid—that a group of people embrace for guiding and evaluating food consumption. Systems of rules and principles, and supporting practices and institutions, might be designed with an eye to furthering the aims of the healthy eating group. Understood as

[50] Rawls, *supra* note 8, at 85-86.

[51] We have benefitted from Bernard Gert, *Morality*, *in* Stanford Encyclopedia of Philosophy, *available at* http://plato.stanford.edu/entries/morality-definition/.

a normative system—and not only a set of "tips" for healthy eating—we would imagine participants praising those who follow the guidelines and criticizing those who do not. Treating the rules and principles of healthy eating as a normative system would be a way of strengthening participants' capacity to resist temptations such as too many rich desserts. Normative systems in society—dedicated to healthful living, etiquette, athletic competition, improved public speaking, or what have you—will often help individuals coordinate so as to advance joint purposes and intentions. But they need not be based in morality, or have much (or anything) to do with morality.

Moral categories of assessment typically also have a special sort of categorical bindingnesss—that is, they are not chosen, and they apply to us regardless of our specific plans and purposes—and they are often thought to be overriding. Each of us is entitled to decide for ourselves whether to regulate our conduct on the basis of the imperatives of healthful eating and, leaving aside our dependents, we do no wrong to others by eating unhealthily. However, we are not entitled in the same way to decide to do without morality: we are required categorically to observe a variety of moral duties to others. These might involve not harming them, or assisting them when they are in dire need and we have a surplus. Precisely which considerations generate moral "oughts" is contested and depends on one's particular conception of morality. A utilitarian will say that we are obliged to act to maximize the net sum of happiness. A contractarian such as Scanlon will say that we ought to treat others on the basis of reasons and principles that they could not reasonably reject.[52] Other-regarding morality is what we owe to other people generally as a matter of imperative duty, whether we like it or not; political morality is what we owe to our fellow citizens.

What about law? H.L.A. Hart and other legal positivists have long argued that legal systems may be understood as normative—as generating standards for assessment and even criticism-- but not necessarily moral.[53] To describe some course of action as legally required does not mean that it is morally required, and vice versa. With that we agree.

An essential feature of a legal system, related to its normativity, is that participants take an "internal attitude" toward the rules and norms that it generates: not simply recognizing that they exist, as would an anthropologist in observing an alien society, but rather recognizing that valid legal rules and principles supply

---

[52] Scanlon, *supra* note 44.

[53] *See* H.L.A. Hart, The Concept of Law (2d ed. 1994); Scott Shapiro, *What is the Internal Viewpoint?*, Fordham L. Rev. 1157 (2006), http://digitalcommons.law.yale.edu/fss_papers/1336.

them with standards for assessment and reasons for action.[54] Indeed, the reasons for action supplied by valid law are "exclusionary," which is to say that with respect to a given issue or policy, the existence of a valid law provides us with a reason that excludes or displaces the reasons we would otherwise have, for example, our own best personal judgment on the issue.

We think that the sort of normativity represented by the "internal attitude" to law, and the exclusionary reasons furnished by valid law, are similar to the attitude that rational actors would have to the shared logic in the positive model. So long as agents conclude that the coordinated equilibrium is better for them than the uncoordinated outcome, the shared logic that serves the instrumental coordinating function suggested by the Hadfield-Weingast model is treated as authoritative and as displacing reasons arising from individuals' personal idiosyncratic logics. And, indeed, rational participants in the community of law would take an internal attitude toward its common logic in evaluating the correct response (punish or not) to potentially wrongful behavior: the relevant question to be asked is no longer what do "I" think, or what does "he" think; it is, what do "we" think, according to our shared logic of coordination based on law? The positive logic of coordination, based as it is on mutual advantage, does not as such give rise to moral obligations. The agents who conclude that they are personally better off abandoning the coordinated outcome are, in this model, free to do so without moral censure.

The positive model addresses only the interests of rational agents in participating in the costly punishment of others who have violated rules, even when the violation does not harm them personally. But the model is easily generalized to take into account the incentive to participate more generally in the sense of observing the rules oneself. Just as my failure to object to a rule violation can undermine the cooperative scheme by generating uncertainty about my ongoing commitment to the scheme, so too can my disregard for compliance with the rules myself. In this more expansive context, it is easy to interpret the employment of a common logic for assessing what is punishable and what is not as an internal attitude towards law. The common logic in this sense provides an individual citizen with reasons for action in accordance with its classifications: by acting in accordance with law, both in terms of personal compliance and an expression of censure for violations by others, citizens communicate that they continue to prefer the coordinated equilibrium to the alternatives and so can be expected to continue to support a particular cooperative scheme.

[54] Hart argues that a legal system may exist where only the officials who operate the system take an internal attitude toward it. *See* Hart, *supra* note 53, at 116-17.

Legal systems are often understood to be normative in the same sense as the positive model: as generating standards of assessment and reasons, not necessarily moral, for action, and as requiring (as a condition of existence) some range of actors to take the "internal point of view" toward the system. We would go further, however, and argue that legal systems are often properly understood as moral in several important respects.[55]

First, important *parts* of legal systems overlap with moral demands: criminal law prohibitions on harm to others, for example, represent both legal and moral requirements. Many legal rights codify and clarify individuals' basic moral claims against one another.

But, second and more broadly, the whole purpose of a system of law is to help a diverse group of individuals to coordinate on common courses of action in conditions of moral diversity. There is no need for law in a simple homogenous world governed by spontaneous norms, but in conditions of diversity, even morally well-disposed individuals need authoritative institutions to clarify and codify what they owe to one another. A sufficient condition for law is disagreement among well-intentioned and conscientious people; as John Finnis says, law "would have a completely adequate rationale in a world of saints. … Intelligence and dedication, skill and commitment … multiply the problems of coordination, by giving the group more possible orientations, commitments, projects "priorities,' and procedures to choose from." [56] Systems of law advance vitally important human purposes and common goods of all sorts by allowing communities to render authoritative judgments and determinations.

So it is hardly surprising that there generally are moral reasons for complying with, and working for the improvement of, decent systems of law that reasonably approximate to justice, even if much of the content of law in such systems is not derived directly from morality. Moreover, as beneficiaries of others' compliance we can have a moral obligation to *do our part in turn* to sustain the system: a duty of reciprocity not to "free ride" on others' actions.

So we think that there is typically significant overlap between law and morality, and that decent systems of law serve vital human interests and purposes. Nevertheless, neither is reducible to the other. We recognize that there is often a gap between legal and moral claims and entitlements: people often have moral claims on others that are not legal claims. And the law may protect forms of conduct that are immoral (refusing assistance to the needy, for example).

---

[55] We have benefitted from the discussion in Andrei Marmor, Philosophy of Law (2011), *see* e*specially* ch. 4.

[56] John Finnis, Natural Law and Natural Rights 232, 269 (1980).

B. Positive Roots for Normativity

The positive model suggests that rational agents with conventional preferences can develop a shared, neutral, and freestanding evaluative system for judging conduct. The logic is normative: it classifies behavior as good or bad, wrongful or not, acceptable or not, punishable or not. This system of classifications is generated, on the positive model, for instrumental purposes: to achieve a set of relationships with others that all judge to make them better off. Participants accept two key facts: the existence of diversity in private evaluative systems and the dependence of each on the others for enforcing compliance with norms that promote valuable social cooperation. In doing so they exploit what some evolutionary theorists call humankind's "cognitive niche": the capacity to employ abstract reasoning and language to achieve better outcomes through communication and cooperation.[57] The common logic is such an enterprise. The common logic and supporting institutions thus, provides otherwise diverse individuals with a shared normative system that is equally available to all.

Individual preferences in the positive model remain those of economic rationality. Participants adopt the internal point of view on law for instrumental purposes: securing the benefits of the coordinated outcome that the agent prefers, on self-interested grounds, to the alternative world without coordination. So, the extension of equal treatment to women or gays or African Americans, or the inclusion of persons with severe handicaps, will depend in the positive model on the projected costs and benefits for interested participants. This makes the stability of the resulting equilibrium still a contingent matter: the content of the common logic is responsive to the particular distributions of power and interests in the population. The positive model predicts that those who are not needed for effective enforcement will be missing from the common logic (indeed, that logic may be averse to their interests). Systems that are stable for shared instrumental reasons may be far from those that are stable for morally defensible reasons.

The movement to genuinely moral standards of assessment depends on participants being willing to take seriously the demands of fairness and other moral entitlements, or the public good as such. Then the inclusion of African Americans, women, or gays will turn not simply on whether this leads to better outcomes for sufficient numbers of individual participants but public convictions concerning just or fair treatment.

But how might a transition from mere rationality to moral reasonableness take place? Adam Smith, in his *Theory of Moral Sentiments,* paints an interesting

---

[57] Steven Pinker, *The Cognitive Niche: Coevolution of Intelligence, Sociality, and Language*, 107 Proc. National Academy Sciences, USA 8993 (2010).

picture of how moral sensibilities might emerge from thinking about how others regard us. Normal humans, Smith claimed, are inherently sociable creatures who desire not only to be thought well of but indeed *to merit being thought well of*. Thus, it is not only disadvantageous but inherently disagreeable to be conscious that one is an object of merited disapprobation in the eyes of others. As creatures with complex mental capacities, Smith suggests, we can form abstract ideas about meritorious forms of conduct: as these become the bases of shared social standards of assessment, we form desires to act according to meritorious criteria, and to be known to do so.[58] This suggests a path from the positive to the moral: from stability for instrumental reasons to stability for the right reasons.

If preferences—desires—are modified in this way, if we cross the bridge from an instrumental normativity that promotes our interests to an authentic morality rooted in concern for others, or rooted in the acknowledgement of moral standards of conduct as such, then principles for inclusion or equal treatment will no longer be contingent on the power of particular groups. Authentic concerns for the welfare of others or the requirements of justice do not depend on perceptions of a benefit for oneself. Moral sensibilities are essential for justifying the inclusion of those who lack the power to significantly assist in decentralized enforcement.

A crucial bridge is crossed, as Smith recognized, when we can imagine participants in systems of social norms reflecting upon the forms of conduct that will win the approval *not only of actual spectators*, but of an *ideal "impartial spectator"* who sees our interactions clearly and fair-mindedly, as would an ideal judge: who judges wisely and well, and not necessarily as the majority of others in our society judge. We on a critical morality when we seek the approval of an ideal spectator. Such ideal-regarding considerations may be introduced to provide a critical perspective on actual or proposed moral rules, or, perhaps, to adjudicate conflicting interpretations of social rules or divergent proposals for revision. In the face of such disagreements the questions arises: what should the rules be? How should power or property or the gains of cooperation be divided?

If I am articulating to my fellows why they should join me in shunning a bad actor, in the language of the common logic, I may claim that a failure to do one's part is not only disadvantageous but unfair: normative public reasoning and language provide me with a currency of reasons for assigning and expressing forms of approval and disapproval that entail potent evaluative attitudes in participants and spectators. The public normative language of assessment furnishes participants with the vehicle for assigning a reputation and circulating it as a matter of common

[58] Adam Smith, The Theory of Moral Sentiments Part III (D.D. Raphael & A.L. Macphie eds., 1976) (1759).

knowledge. It also furnishes grounds for self-assessment: for shame or pride depending on one's behavior.

These observations are strengthened by recent work in evolutionary theory and behavioral game theory suggesting that human beings are frequently prepared to reciprocate acts of cooperation even when doing so is personally costly. Likewise, with no expectation of personal gain, they are frequently prepared to engage in costly acts of punishment of those who fail to cooperate or do their share. Other evidence suggests that individuals tend to be "inequality averse": willing to bear some cost to increase equality in the group, and even willing to bear additional costs to "reduce the payoffs of relatively favored individuals even more."[59] Herbert Gintis argues that the evidence strongly suggests that people are not primarily motivated to be rational maximizers of their own interests, but rather "conditional altruistic cooperators" who are disposed to value fair cooperation.[60] Bowles and Gintis argue that humans have social preferences: "a concern for the well-being of others and a desire to uphold ethical norms" and virtues of character.[61]

The citizens in the stable equilibrium posited by Hadfield and Weingast, we suggest, can be expected over time to enlist the assistance of praise and blame, reputation, and then conscience in a manner similar to that described by Smith: we judge others and are conscious of being judged ourselves, and so we come naturally to inspect ourselves as if by others. The progression to an ethics of critical moral judgment represents a further shift in motivation. Those who are capable of critical morality, as represented by Smith's "impartial spectator," take a critical attitude toward social categories and public opinion, and recognize the difference between what is praised in one's society and what is truly *praiseworthy*.[62] Moral judgment does not simply register the voice of others or the dictates of individual or collective advantage. The public moral judgments reflected in public reason are similar: proposed principles of social justice must be tested against the strongest available competitor moral and political theories to discern which provides the strongest candidate principles—on the moral merits—for describing what we owe one another as equal participants in a shared social and political order. On the liberal account, we seek to formulate principles of social justice that can gain widespread assent but that are also sound (and acceptable) based on our best understanding

---

[59] Herbert Gintis, The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences 48 (2009).

[60] *Id*. at 56-75.

[61] Samuel Bowles & Herbert Gintis, A Cooperative Species: Human Reciprocity and Its Evolution 10 (2011).

[62] *Id*. at 8-18.

of what we owe to one another as moral agents: as cooperators like us capable of acting on fair terms of cooperation.

C. The Stability of Publicly Reasoned Political Morality

We want to say a final word about the nature of the stability that may be possible in a world that is populated by agents that are not only rational and rationally (or prudentially) reasonable, but also moral. Moral agents in Rawls's sense support the political conception of justice because, and to the extent that, it reflects a fair system of cooperation, and answers to other principles of political morality. In the Hadfield-Weingast model, such agents act collectively to punish those who transgress shared terms of cooperation as a means of confirming on an ongoing basis their commitment to the normative categories of the common logic. As a positive matter, the movement from a purely positive logic to a system of shared reasons and decentralized enforcement that is understood to be normative by participants represents not only a transition in the nature of stability—toward stability for shared moral reasons—but also a deepening of the grounds of stability, and a more robust equilibrium. When agents' preferences are transformed, such that they internalize moral (and not merely socially normative) standards and thereby come to desire to act fairly in a regime of reciprocal fairness, the costs and benefits of compliance and enforcement are similarly transformed. Voluntary compliance will then gain further support from consciousness of participation in a moral community. This reduces the need for social punishment and the losses inevitably associated with detection and sanctioning. Even where social sanctioning continues to play a role, the costs of punishing transgressions is reduced, diminished by the positive benefit one derives from acting in a manner consistent with one's endorsement of the public moral system. Note too that the costs of being punished rise when one is the object of moral disapproval: one is then conscious of others' moral disapproval, and this is bound to be especially painful to normally reflective agents, as Smith observed.

When one's motives for compliance and also enforcement are rooted in one's moral judgments and evaluative principles, and not only in one's interest-based calculations, they are much more deeply rooted in and integral to our entire evaluative system: at a more profound level, or in the way we see ourselves and the world around us. This could of course have several downsides: flexibility might be reduced and intolerance could become more of a danger. Insofar as shared norms help mark the boundaries of particular groups, the moralization of norms may go along with group mobilization for violence.[63] In the face of deeply moralized

[63] See the important account in Russell Hardin, One for All: The Logic of Group Conflict (1995).

religious intolerance, social observers extol the cooling effects of interest-based calculations.[64] We admit these dangers of excessive moral zeal and severity. The presence in society of a great diversity of religious and philosophical world-views may be helpful insofar as individuals thereby recognize that their public moral commitments are not the whole of their beliefs about what is right and wrong, good and evil.

Moreover, the public moral judgments that furnish a basis for social cooperation must be ones that have broad appeal and that are fairly easily understood. They are liable to be, or become, aspects of the shared public culture: part of the common currency of mutual assessment and accountability. Our consciousness of their being both moral and shared contribute to the legitimacy and salience of the social rules that are based upon them. The shared logic of cooperation is now no longer simply an equilibrium based on mutual advantage (as in the rational model of Hadfield and Weingast), but rather has the force of collective moral assessment.

However, there are also features of the liberal democratic normative model that specifically help guard against moralized intolerance. First of all, there is the priority given to liberty: citizens in liberal democracies are frequently highly sensitive to improper restraints on expression, even when these are merely normative rather than legal. Moreover, citizens who agree on a shared logic of coordination nevertheless continue to espouse a wide array of diverse personal evaluative systems, including differing religious and philosophical views. Both models (positive and moral) regard the shared system of reasons and principles as freestanding and complete on its own terms—as a distinctive autonomous and integral system—but *also* as open, in an appropriate way, to criticism and input from citizens' differing idiosyncratic non-public systems of belief, that is, from complaints or objections arising from people's religious views or ethical ideals. Full public deliberation must include moments when citizens assess the shared principles from the personal standpoint of their beliefs as a whole. Each of us must be able to assure ourselves that we can live, in good conscience, with the deliverances of public reasons.

This duality of normative commitments is an important critical check on any conclusions that are reached by the shared public logic. The wide diversity of normative communities – religious and other associations and groups of all sorts – will nurture critical perspectives that will frequently contend with one another and challenge reigning convictions. Given a healthy set of public institutions and political culture, it can be hoped that the "arc of history" bends toward greater justice.[65]

---

[64] For a wonderful account, see Albert O. Hirschman, The Passions and the Interests: Political Arguments for Capitalism before Its Triumph (1997).

[65] To borrow from a valuable book (and essay) by Joshua Cohen, The Arc of the Moral Universe and other Essays (2010).

V. Conclusion

This long argument has tried to display some ways in which public reason can be seen to play an important and constitutive role in the path from a *modus vivendi,* to a more robust form of stability for the right reasons. The Hadfield-Weingast model argues that purely self-interested rational agents engaged in various forms of social cooperation will develop the institution of a neutral common logic that helps them to coordinate the enforcement of norms that promote their individual well-being. Self-interested but enlightened agents who can see beyond the immediate present can recognize the value of stable forms of cooperation guided by a shared logic of reasons and principles that all can share as equals while differing about their personal evaluative standards or conceptions of the good.

A shared system of public reasons thus constitutes a shared fund of common meanings that agents engaged in cooperation can express to one another to signal their mutual commitment to cooperation on shared mutually intelligible terms. Shared, public reasons solve a problem of mutual assurance as a matter of common knowledge, making it rational for agents to participate in joint enforcement. This seems to us very much the sort of work that public reason understood as a public morality also performs, save that the agents now share a *morally* evaluative language of fair cooperation. The addition of this normative dimension makes sense from a positive standpoint: now participants in social cooperation have available a richer and more powerful set of categories for evaluating one another's, and one's own, conduct and expressing either approval or disapproval. Our capacity to mutually assess and sanction and reward is thereby deepened, and the bonds of social cooperation ought thereby to be strengthened.

Without collapsing the categorical difference between positive and normative analysis, we think these two perspectives prove to be mutually complementary here. Positive analysis strengthens the case for public reason understood in normative terms. And the development of standards of moral assessment makes sense from a positive point of view.