2012

What is Law? A Coordination Model of the Characteristics of Legal Order

Gillian K Hadfield, *University of Southern California Law*
Barry R. Weingast, *Stanford University*

Journal of Legal Analysis Advance Access published July 29, 2012

WHAT IS LAW? A COORDINATION MODEL OF THE CHARACTERISTICS OF LEGAL ORDER

Gillian K. Hadfield, and Barry R. Weingast[1]

ABSTRACT

Legal philosophers have long debated the question, what is law? But few in social science have attempted to explain the phenomenon of legal order. In this article, we build a rational choice model of legal order in an environment that relies exclusively on decentralized enforcement, such as we find in human societies prior to the emergence of the nation state and in many modern settings. We demonstrate that we can support an equilibrium in which wrongful behavior is effectively deterred by exclusively decentralized enforcement, specifically collective punishment. Equilibrium is achieved by an institution that supplies a common logic for classifying behavior as wrongful or not. We argue that several features ordinarily associated with legal order—such as generality, impersonality, open process, and stability—can be explained by the incentive and coordination problems facing collective punishment.

1. INTRODUCTION

What is law? What distinguishes legal order from spontaneous social order? How can we identify when a community is governed by the rule of law? What institutions support the effort to pattern behavior on the basis of deliberately chosen legal rules?

1 Hadfield, University of Southern California, Gould School of Law, Los Angeles, CA 90089 USA, E-mail: ghadfield@law.usc.edu; Weingast, Stanford University.
We have benefitted from helpful comments and suggestions from numerous colleagues. For particularly detailed suggestions and advice we are grateful to Kenneth Arrow, Sam Bowles, Ryan Bubb, Randall Calvert, Chuck Cameron, Bob Cooter, Mariano-Florentino Cuellar, Giuseppe Dari-Mattiaci, John Ferejohn, Bob Gibbons, Les Green, Carmine Guerriero, Lewis Kornhauser, Antoine Lallour, Shmuel Leshem, Steve Macedo, Andrei Marmor, Richard McAdams, Giorgio Monti, Paul Roemer, Philip Pettit, Josh Ober, Katharina Pistor, Edward Stiglitz, Eric Talley, Brian Tamanaha, Mike Tomz, Joel Trachtman, and John Wallis. We have also benefitted from discussions at workshops at the National Academy of Sciences Sackler Colloquium, the Center for Advanced Study in the Behavioral Sciences, the American Law and Economics Association Annual Meeting, the Annual Meeting of the International Society of New Instituitional Economics, the Comparative Law and Economics Forum, New York University, the University of Toronto, the University of Texas, the University of Paris, the University of Amsterdam, and Harvard University.

doi:10.1093/jla/las008

These questions lie at the heart of numerous projects in economics and politics—explaining the evolution of social order in human communities, building markets to support economic growth in poor and developing countries, establishing the necessary architecture for stable democratic governance, managing the increasingly integrated transactions of a globalized web-based economy. Nonetheless, economists and positive political theorists to date have had almost nothing to say about these questions.[2] Most work in economics and positive political theory (PPT) simply presumes that legal order is defined by the existence of the institutions that characterize modern western democracies; namely, centralized production of legal rules by legislatures and courts combined with centralized coercive enforcement of those rules by duly constituted governments. The vast majority of economic and positive political theory focuses on the substance of legal rules but not the characteristics of distinctively legal order per se.

In this article, we initiate the project of filling the gap in law and economics and PPT by developing a rational choice model of legal order. The principal goal of our work is to develop an account of legal order that does not presume that legal order is characterized, necessarily, by the types of institutions we see today in modern developed nation states: courts, legislatures, police, and so on. Legal order arises in so many different environments, ranging from early human societies prior to the development of the nation state to a globalized interdependent civil society that in many ways transcends the nation state. Developing a systematic social scientific account of law that identifies the conditions under which legal order emerges and is stabilized requires that we abstract from the particular institutions embodying law in modern nation states.

In particular, we claim that it is critical that a social scientific account of law does not presume that law is necessarily characterized by centralized punishment—delivered by a formal institution with coercive power, such as a government. Our framework thus allows for the possibility that law is enforced by decentralized mechanisms and the model we present in this article demonstrates that we can achieve legal order exclusively on the basis of decentralized enforcement, without any centralized coercive authority. This possibility marks a major departure from the implicit definition of law employed in most economic and positive political theory. As Dixit (2006, 3) observes, "conventional economic theory... assumes that the state has a monopoly over the use of coercion". Ellickson (1994, 127) defines law as rules that are enforced by governments rather than social forces.

2 As we discuss in more detail below, Kornhauser (2004) is a rare exception.

By decentralized enforcement, we mean that the imposition of penalties results from individual decisionmaking among ordinary agents acting independently, not the decisionmaking of official legal actors such as police or judges nor the result of express pacts for collective action. Decentralized enforcement may also include voluntary compliance (in the sense that the individual "punishes" himself or herself for engaging in wrongful conduct), individual punishment (as occurs when someone plays a tit-for-tat strategy in a repeated game [Axelrod 1984], for example), or collective punishment (as when a set of individuals, acting independently, collectively refuse to deal with someone who has done something wrong). As we discuss in a companion paper (Hadfield & Weingast 2011a), a wide range of examples exist of settings in which legal order is apparently achieved without the existence of a centralized coercive enforcement body; including, medieval Iceland, Gold Rush California, medieval Europe under the Law Merchant and merchant guilds, and modern international trading and collaboration regimes.

Our reason for excluding the existence of a centralized enforcement body from our model is to develop a framework capable of analyzing a range of questions concerning if and when centralized enforcement of law is necessary or sufficient to secure legal order. These are critical questions if we seek to explain the emergence of legal order prior to the organization of states with a monopoly over legitimate force, the potential for establishing the rule of law in environments with weak or corrupt governments, or the feasibility of establishing legal order in exclusive reliance on centralized coercive force (i.e., a system that ignores the role of decentralized mechanisms in structuring legal order). Analysis of these questions is not possible if we follow the dominant assumption in economics and political science that law is, by definition, a set of rules enforced by government.

We therefore propose a different definition of legal order to guide the development of a positive framework for analyzing the emergence and characteristics of law. We will say that an environment can be said to be organized on the basis of legal order if (i) there is an identifiable entity (an institution) that deliberately supplies a normative classification scheme that designates some actions as "wrongful" (punishable, undesirable) and (ii) actors, as a consequence of the classification scheme, forego wrongful actions to a significant extent. We propose these criteria as indicia of legal order for the purposes of positive analysis because they capture the fundamental policy role of law that makes the study of law of interest to economists and political theorists: law is a purposeful vehicle for shaping behavior to achieve desired ends. Many mechanisms and institutions shape behavior, of course: social norms guide conduct in all human societies, for example. What makes law of interest to economists and positive political theorists, however, is the capacity for deliberately changing behavior patterns in order

to achieve normative goals such as economic growth, the security of individual freedoms, or the redistribution of wealth. Our proposed criteria for identifying legal order aim to distinguish a system in which norms are emergent, arising as a matter of practice from repeated interactions, and one in which norms are deliberately articulated and systematically implemented.

Our approach thus tracks a basic distinction drawn by H.L.A. Hart (1961/1997): between a regime in which there are only primary rules of behavior, and one in which there are also secondary rules that can introduce or change primary rules. Hart called the system with both primary and secondary rules "law". As Hart recognized, a system without secondary rules is dependent on slow adaptation to respond to changes in the environment or desired outcomes; a system with secondary rules is capable of deliberate effort to change behavior.

As with Hart's definition, our definition has to be understood as minimal: there may well be other systems of achieving behavioral order that could be said to be characterized by secondary rules or an identifiable entity that deliberately supplies a normative classification scheme. We leave open, then, the challenging task of fully distinguishing legal order from other forms of deliberate social order. (For further discussion of this point, see Section 3, below.)

In this paper, we take up instead the more focused task of showing that an equilibrium that meets our minimal criteria for a legal order can be secured in a setting in which penalties for wrongful behavior are delivered exclusively by decentralized collective punishment. That is, we demonstrate the possibility of legal order without a third-party centralized enforcement mechanism, such as the state or some other powerful actor with an effective monopoly of force. Moreover, we demonstrate that the equilibrium displays several attributes that are often associated with our intuitions about the nature of law or the rule of law. The legal order in our model is characterized, for example, by the existence of general rules and impersonal abstract reasoning implemented by open, public, and neutral procedures. These are attributes that many legal philosophers (e.g., Fuller 1964; Raz 1977) associate with the concept of law or the rule of law. (We discuss this literature in more depth in Sections 3 and 4.) In our model, these attributes are directly attributable to sustaining the efficacy of decentralized collective punishment. This is in contrast to the conventional focus in legal theory on the relationship between the attributes of law and the capacity of an individual to be guided by rules or the normative limits on the exercise of force by a coercive power such as a government.

The challenge of sustaining decentralized collective punishment is the challenge of coordinating individual decisions to participate in delivering costly penalties to those who engage in wrongful conduct. Our model therefore presumes that effective punishment—which deters wrongful conduct in equilibrium—requires coordination among multiple agents who must make

simultaneous decisions about whether to punish a wrongdoer. Clearly, such coordination requires that a sufficient number of agents identify a particular act as wrongful. We demonstrate that this coordinating function can be served by what we call a *common logic*, a system of reasoning that generates unique common knowledge classifications of conduct. We argue that such common knowledge classifications can be provided by a third-party institution that supplies a system of public and impersonal reasoning by exercising what we call *authoritative stewardship*.

We do not assume, however, that there is an inherent incentive to participate in collective punishment based on these common knowledge classifications. That is, we do not presume that coordination is sufficient to support an equilibrium, as does the existing literature analyzing the role of coordination and convention in law. Because punishment is costly to individuals, we must explain why individuals participate in a system that relies on collective punishment. In this regard, we track a growing literature on collective punishment that investigates the puzzle of why people, in many cultures (Henrich et al. 2006) and in experimental settings (Fehr & Gachter 2002) are willing to incur positive costs to inflict penalties on those who behave wrongfully without any immediate material benefit. (We canvas the literatures on coordination and collective punishment in more detail in Section 4.)

We link the resolution of the incentive problem to the characteristics of the coordinating institution—that is, to the attributes of a legal order. The incentive to punish that we identify is the incentive to alter beliefs—held by those who may engage in wrongful conduct and those who might participate in punishment—about the likelihood that wrongful acts will be met with an effective punishment. More precisely, an individual punishes in order to signal to other agents that an equilibrium with punishment based on the common logic is or continues to be in the individual's private interest. In our model, the participation of all agents is necessary for effective punishment. The failure by any individual to carry through on punishment in the event of wrongful conduct leads other agents to infer that collective punishment is no longer sustainable. This inference destroys both the incentive of others to punish and the deterrence of wrongful conduct.

We show that equilibrium with effective collective punishment then depends on the generality, stability, openness, and impersonality of the common logic. These attributes secure, in our model, the incentive of individual agents to participate in collective punishment. Universality and impersonality—in the particular sense of being addressed to the interests of all and independent of the reasoning of a particular entity—ensure that an individual can expect that a system of collective punishment will be of personal benefit. Stability ensures that today's decision to punish based on a common logic conveys information about future benefits under that same logic. Openness assures heterogeneous

individuals, who possess what we call an *idiosyncratic logic* for classifying wrongs against them, that they will have access to a mechanism for integrating their personal classifications into the common logic.

An important implication of our model is that it provides a link between the attributes of legal order that many intuitively associate with law and the resolution of the coordination and incentive problems that underpin effective collective punishment. This raises a question of whether a legal system that relied exclusively on centralized punishment to deliver penalties, as the great majority of work in economics and PPT assumes, would display the attributes generally associated with the rule of law.

Our framework, therefore, makes a contribution to three literatures. Law and economics, and PPT and the law, both fail to explain why law has various legal attributes, such as those identified in the legal philosophy literature. Law and economics typically defines law exclusively in terms of its capacity to coercively enforce a result. Indeed, law and economics treats law as another form of regulation; that is, constraints enforced by the government. PPT and the law largely treats law as just another source of politics and policymaking. Work in both fields focuses on efficiency, distribution, or other policy characteristics of the outcomes generated by legal institutions. Neither field attempts to explain why legal institutions possess the attributes that mark them as legal. Conversely, although legal philosophers focus on identifying the attributes that distinguish legal systems from other normative systems, they do not address the positive question of how a system of law characterized by these attributes can be generated or sustained. Our approach addresses all these issues.

Our article is organized as follows. Section 2 presents the model and Section 3 extracts the attributes of the equilibrium institution that generates legal order in that model. In Section 4, we relate our approach to the existing literature. Section 5 provides some concluding observations.

2. MODEL

Our goal in this model is to demonstrate that a third-party institution that supplies a common logic for classifying behavior can, if the institution possesses certain characteristics, effectively coordinate decentralized enforcement efforts to deter behavior that reduces social welfare. We therefore model a setting in which other means of deterring welfare-reducing behavior—such as changes in organizational or transactional terms or unilateral or centralized punishment—are either unavailable or exhausted. This sets up our problem of determining if it is possible to deter the residual welfare-reducing behavior with decentralized collective punishment alone.

We first provide an overview of the model and our modeling strategy, and then put the model into formal terms.

2.1. Overview

The setting we consider is one in which there is an actor, *S*, engaged in repeat and independent interactions with two individuals, *A* and *B*. In each interaction with *A*, there is some probability that *S* has an opportunity to take an action that benefits *S* but may harm *A*. Whether the action harms *A* depends on *A*'s private reasoning. *B* faces the same risk in each period as *A*: suffering what it judges privately to be a reduction in welfare because of an action *S* has taken. *A* and *B* engage in independent relationships with *S*: an action that *S* takes in its relationship with *A* has no impact on *B* and vice versa. Nor do *A* and *B* care about what happens to the other; they do not hold preferences over how the other is treated. The only available means of deterring *S* is a decentralized collective punishment, in which *A* and *B* make simultaneous decisions to incur some personal cost to inflict loss on *S* when one of them suffers a loss. Delivering this punishment, therefore, requires that both are willing to act what appears to be altruistically: punishing *S* for harming someone else, despite the absence of pro-social preferences.

We make this setup concrete by assuming that *S* is a seller of goods and *A* and *B* are buyers.[3] The actions *S* can take in its relationships with a buyer are cost-cutting measures. Some of these cost-cutting measures will be judged by the buyer, according to the buyer's private system of classification—what we call the buyer's *idiosyncratic logic*—to be wrongful. In the contracting setting, this means that the buyer judges that the seller's actions reduce the buyer's expected profits below the level that the buyer believes to be promised by the contract.

We do not make any *a priori* assumption about whether the buyer's classification of actions as wrongful is objectively valid. A key feature of this model is that we pay close attention to the potential for a diversity of views about what it means for the seller's performance to be wrongful. This diversity of views makes

3 Although we explicate the model in concrete terms by describing a sale transaction, nothing in the model is particular to this setting. The model can be interpreted as applying to any setting in which there is a potential wrongdoer who may exploit a community of potential victims who have the capacity to impose some penalty on the wrongdoer. Our seller, for example, could be a feudal lord and our buyers, serfs. Serfs then invest upfront in working the land, at the risk that in the future the lord will extract a wrongful share of the harvest or wrongfully impose additional duties or conditions or hardships. Alternatively, we could interpret the seller as a powerful private individual and our buyers as other private individuals who are at risk of physicial aggression or theft at the hands of the powerful individual. The potential victims make investments each period in building up property, for example, or operating in public spaces, at risk that doing so gives the powerful individual an opportunity for theft or aggression.

the determination of what is wrongful inherently ambiguous. Moreover, we assume that each buyer's idiosyncratic logic is fully private and inaccessible to others at reasonable cost. This rules out the possibility that the various actors can simply deduce how individual buyers will classify conduct. By designating individual classification schemes as "idiosyncratic", we emphasize that the classification is conditional on the buyer's particular circumstances and that the buyer's decision to buy is based on its own evaluation of the circumstances in which the deal is valuable. We posit idiosyncrasy not as a form of odd or unusual preferences but rather as a source of value-generating diversity in an economy (Hong and Page 2001).[4] Idiosyncratic assessment is particularly likely to arise in dynamic environments where individuals are discovering and inventing new possibilities for transactions. This is true over significant periods of human economic history. For example, a buyer may discover ways to reduce production costs by using just-in-time inventory in an environment in which all others use conventional inventory practices. This innovation will lead the innovating buyer to assess the value of delayed deliveries differently than a conventional buyer and, at least initially, this assessment will be unknown and perhaps inscrutable to others. We do not model the buyer's logic as private information, although it may depend on private information. Rather, the buyer's logic is a system of private reasoning to organize and analyze information and to make judgments about how to classify behavior.[5]

In addressing our questions, we drastically simplify the relationship with and the role of the seller: we assume the price at which the buyer and seller transact is fixed and that the seller's only decisions concern the choice of how to perform the contract with each buyer in each period. We recognize that there are transactional adjustments that the parties can and are likely to make as the efficacy of deterrence shifts, but so long as any such adjustments nonetheless leave residual risks that the seller will engage in behavior that the buyers individually would prefer to deter, our results continue to hold. We emphasize that our goal is not to predict the attributes of the transactions between buyer and seller but rather

4 Hong and Page (2001) present a model in which "collections of agents outperform individuals partially because people see and think about the problems differently" (p. 130). Diversity is captured by characterizing individuals in terms of their individual *internal language* (used to represent objects), *perspective* (a mapping from objects into the internal language) and a *heuristic* (a set of rules for moving around the space of objects in his or her internal language, a logic).

5 Crawford & Haller (1990) present the idea that agents may lack a common language for representing the structure of a game and thus cannot reproduce the reasoning of others (for purposes of coordination) except on the basis of observed outcomes that can be uniquely associated with a particular action. See also Kramarz (1996) solving an N-player coordination game in the absence of a pre-existing common language. Both Crawford & Haller and Kramarz analyze the dynamic process of reaching coordination through the generation of a common language based on the evolving history of a game.

the attributes of a mechanism by which the buyers can secure some deterrence through collective punishment in the absence of a centralized enforcement authority.

We focus on collective punishment because in many interesting settings unilateral punishment is ineffective. More generally, collective punishment will often be less costly and/or generate greater benefits than unilateral punishment. A two-buyer one-period boycott, for example, inflicts a higher cost on the seller than a one-buyer two-period boycott (because the 2nd period loss is discounted by the seller). Moreover, a threat of collective boycott can deter (as we will show) wrongful performances vis-à-vis both buyers; a lone boycotter's threat can deter at best wrongful conduct vis-à-vis that buyer. In a more general model, in which a particular buyer does not interact every period with the seller, a collective punishment scheme would call for a penalty to be delivered by buyers who interact sooner with the seller and thus can deliver a more immediate and hence more costly punishment. We discuss examples of historical and contemporary settings in which legal order depends on collective punishment in Hadfield and Weingast (2011a,b). Theorists in evolutionary game theory (see, eg., Boyd, Gintis & Bowles 2010), behavioral economics (see, e.g., Fehr and Gachter 2002; Henrich et al. 2006) and institutional economics (see, e.g., Milgrom, North, & Weingast 1990) have also emphasized the importance of collective punishment schemes in the evolution and diversity of social order. (For further discussion of this literature, see Section 4.)

We show that it is possible for an institution that provides a public classification scheme (but that lacks any enforcement powers) to coordinate and incentivize the buyers' participation in collective punishment that achieves effective deterrence of the seller in equilibrium. We call the institution that provides this classification scheme a *common logic*. Examples of institutions supplying a common logic include "English common law", "the Law Merchant", "the customs of this village as articulated by the elders", "rabbinical teachings", "the Dutch merchants' guild", "the Archbishop of Hamburg", and "the United States Supreme Court". We emphasize that a logic is an institution, not a disembodied classification scheme. The classifications reached by a logic depend in part on the procedures used to implement classification. For example, the English common law includes a set of rules about what counts as valid evidence and who may present arguments to persuade a group of individuals known as judges about how to classify performance; the elders of a particular village are likely to have a different set of characteristics regarding process, evidence, and so on that affect the classification system.

We show that a common logic capable of coordinating and incentivizing collective decentralized punishment to deter wrongful seller behavior possesses characteristics that are usefully identified as characteristic of legal order.

We do not model the process by which legal order emerges or by which legal order can be created. Our focus in this article is only on the characteristics of the equilibrium legal order we model. The proof we offer is in the form of an existence proof. We leave the question of the conditions under which legal order can be predicted to emerge to future work. We turn now to the formal setup of the model.

2.2. Formal Model

Assume there are two infinitely lived buyers, A and B who in each period $t = 1, \ldots, \infty$ decide whether to purchase a good from an infinitely lived seller, S. Future profits are discounted with a common discount factor, δ. Buyers value the good at V and contract with the seller to pay a price $P < V$ prior to delivery. The seller incurs a cost, c, to perform on the contract and deliver the good as promised. Let the seller's performance in period t be characterized by an vector $X = (t, x_1, x_2, \ldots x_n)$ of factors with $x_1 \in \{A, B\}$ indicating the identity of the buyer. The elements of X capture a wide variety of considerations relevant to the buyer's and the seller's assessment of the value of a period t deal: attributes of the seller and the buyer, the buyer's use for the good, discussions and correspondence at the time of contracting, promised delivery date, the history and nature of the relationship, the type and quality of good, the location of delivery, location of production, delivery method, insurance terms, risks of loss or damage including "force majeure" type risks, etc. We will write X_t when it is important to identify the time period in which a performance occurs. For each buyer i, let $\mathfrak{X}^i$ represent the set of n-tuples X^i in which the identity of the buyer $x_1 = i$.

In each period, with some probability, the seller has an opportunity to choose a particular performance X that will save the cost c vis-à-vis one of the buyers. For example, in one period the seller might realize an opportunity to change delivery services; in another period the seller might have an opportunity to change packaging or raw materials. In some cases, a cost-cutting action by the seller will reduce the value of the performance to the buyer. Only one such opportunity arises each period; the seller does not face the opportunity to cut costs vis-à-vis both buyers simultaneously. For simplicity, we assume that the seller's choice either preserves the full value the buyer expected from the contract, V, or reduces value to 0. A performance that the buyer judges to reduce value is deemed *wrongful* by the buyer. We model the buyer's judgment in terms of the buyer's idiosyncratic logic, I^i, which maps the (potentially very large) set of possible performance vectors X^i into a binary classification, wrongful (0) and honorable (1):

$$I^i : \mathfrak{X}^i \rightarrow \{0, 1\}.$$

Let θ^i be the unconditional probability that the seller will realize an opportunity to cut costs by engaging in a performance that buyer i judges to be wrongful. The seller's performances are observable to all. $I^i(X^i)$ and θ^i are not discoverable at reasonable cost by the seller or buyer j. The seller keeps the payment P regardless of whether performance is judged wrongful or not by the buyer.

We assume that both buyers will purchase the good in period t even if they anticipate that the seller will exploit all opportunities for cost-cutting, including those that result in a performance that a buyer judges to be wrongful. Expected per-period profits for a buyer i in this case are

$$(1 - \theta^i)V - P > 0.$$

Both buyers will therefore purchase despite the risk of what they judge to be wrongful performance if

$$\theta^i < \frac{V - P}{V}, \quad i = A, B. \tag{1}$$

We assume this condition is met.

We assume an institutional environment as follows. No third-party institution exists that is capable of enforcing a penalty against the seller for a wrongful performance. This assumption rules out, for example, a state enforcement agency. It also rules out the capacity for the buyers to join forces in a private organization that they fund for the imposition of penalties. Any penalties imposed on the seller must therefore be imposed by the buyers acting independently. Thus, the environment is one in which only decentralized enforcement is possible.

A buyer acting independently might inflict on a seller a wide variety of losses in response to wrongful conduct. For example, a buyer might inflict reputational harm on the seller or engage in physical retaliation. We model the penalty available to buyers as a boycott: depriving the seller of profits. The insights of our model, however, do not depend on this form of penalty; any penalty that is costly to both the buyers and the seller and that the buyers can inflict independently will serve the purpose.

We assume that the seller will be deterred from choosing a performance X_t when the opportunity for cost-cutting arises if the seller expects that X_t will be met by a simultaneous boycott by both buyers in period $t+1$. Formally, the seller who avoids wrongful performance for both buyers anticipates a two-period profit of

$$2(P - c) + 2\delta(P - c). \tag{2}$$

The two-period profit for the seller who engages in a wrongful performance vis-à-vis one buyer and anticipates that both buyers will therefore boycott in the next period (yielding the seller a profit of zero in the 2nd period) is

$$P + (P - c). \tag{3}$$

We assume

$$c < \frac{2\delta}{1 + 2\delta} P. \tag{4}$$

This assumption gives us that

$$2(P - c) + 2\delta(P - c) > P + (P - c)$$

so that the threat of a two-buyer one-period boycott deters the seller from taking advantage of an opportunity to cut costs by engaging in a wrongful performance vis-à-vis one of the buyers. We call this a threat of an *effective boycott*.

We then ask the question, is there an equilibrium in which a two-buyer one-period boycott serves as a threat that can increase both buyers' payoffs by deterring the performances each judges to be wrongful?

2.2.1. Coordinating boycotts

To understand what is required to generate an effective two-buyer one-period coordinated boycott that deters wrongful performance in equilibrium, consider first the following strategy, call it strategy *I*, for the buyers:

Purchase good in period t if and only if either
$I^i(X^i_{t-1}) = 1, \quad i = A, B, \quad$ or
no purchases were made at $t - 1$

Strategy *I* calls for a boycott when either buyer judges the performance it received to be wrongful. If a seller expects the buyers to play this strategy, the seller's best response is to avoid wrongful performance. These strategies can therefore support a Nash equilibrium in which the seller adopts the strategy of never selecting a pair of performances (X^A, X^B) such that either $I^A(X^A) = 0$ or $I^B(X^B) = 0$. This is an application of the Folk Theorem for repeated games (Fudgenberg and Maskin 1986).

There are two reasons why we should not expect this equilibrium to emerge in practice, however. First, on a theoretical level, Nash equilibrium is not an appealing equilibrium concept because it is not subgame perfect: put more simply, the buyers' threat to boycott is not credible if the seller in fact engages

in a wrongful performance. Once the seller has cheated in this way, the loss to the cheated buyer is effectively a sunk cost. Boycotting is costly to the buyers (who forego the preferred payoff of purchasing even on the assumption that the seller will choose a wrongful performance whenever an opportunity arises, see Equation (1)). And boycotting does not in fact change the seller's expected payoffs after the boycott ends, so it does not alter the probability of a buyer being cheated in the future; the seller's payoff was only affected by the threat of a boycott. The buyers are better off responding to the wrongful performance by absorbing the loss and continuing to purchase as before. Formally, because the seller is not better off in the equilibrium of the repeated game with deterrence of wrongful performances (this follows from our assumptions that buyers strictly prefer to buy even if the seller cannot be deterred and that prices are fixed[6]), the seller cannot be incentivized to participate in strategies that punish buyers for failing to punish wrongful performance when punishment is called for in the equilibrium strategy.[7]

We are interested in a second reason for thinking that the proposed boycotting strategy above will not support an equilibrium: as a practical matter, even if it was judged to be credible, the boycotting threat is not implementable. This is because the seller and buyer *j* do not know, and cannot at reasonable cost learn, $I^i(X^i)$, buyer *i*'s idiosyncratic classification of alternative performances.[8] As a result, buyer *j* cannot condition its boycotting behavior on what buyer *i* judges to be wrongful. Nor can the seller condition its choice of how to respond to an opportunity for cost-cutting on the potential for the choice to trigger a boycott from the buyers. This latter constraint also rules out the possibility that the

6 Our restriction on price adjustment is intended to be a device to generate a setting in which transactional terms alone cannot generate perfect performance in the repeated game. Clearly such settings exist in practice, or else we would have no need of third-party punishment to deter some wrongdoing.

7 The generalizations of the Folk Theorem that show that a subgame perfect equilibrium can be achieved (Fudenberg and Maskin 1986) require that players who fail to punish are themselves punished.

8 This does not mean that a buyer cannot convey any of its requirements for adequate performance to the seller. We are focusing on the settings in which there is a problem of deterrence; in the contracting context, this refers to settings where it is hard to express and communicate all of the expected attributes of performance in concrete terms up front. Idiosyncrasy preserves the problem we are interested in, which we expect to arise in settings with diverse and innovative economic relationships. A seller may interpret a buyer's emphasis on precisely on-time delivery, for example, as mere exhortation in a context in which all other buyers are satisfied with delivery within a few days of a contract target date. The model can be interpreted in this particular context as one of incomplete contracting, where incompleteness is driven by the difficulty of articulating or predicting ex ante the wide variety of future conditions that can affect the value of performance. For more discussion of this in the context of contracting and innovation, see Bozovic and Hadfield (2012).

buyers could simply announce when they judge they have received a wrongful performance: the seller cannot be deterred if the seller cannot predict those announcements and this lack of predictabilty of $I^i(X^i)$ to actors other than buyer i implies.

This objection to an equilibrium supported by strategy I is a more general critique of the literature on relational contracting and informal enforcement mechanisms. Most game-theoretic models of reputation and coordinated punishment—of which Milgrom, North & Weingast (1990) is an example—assume a unique common classification of actions as "cheating" or not. In fact, what constitutes "cheating" will often be difficult to determine and vary from agent to agent. This is the reason we begin with the assumption that our agents employ idiosyncratic logic to decide what, for each of them individually, constitutes "cheating". This then points up the essential challenge of enforcement in the absence of a third-party coercive force.

2.2.2. *Common Logic:* R

We propose a solution to the twin problems of securing credible threats of punishment and overcoming diverse beliefs about what constitutes "cheating". To do so, we posit the existence of a third-party institution supplying a common logic, R, that is capable of classifying performances involving either A or B. Formally,

$$R: \mathfrak{X}^A \cup \mathfrak{X}^B \rightarrow \{0, 1\}.$$

R is assumed to be inaccessible to the buyers and the seller. We suppose, however, that in each period, the institution can make a public and common knowledge representation of R, which we denote $\hat{R}_t$:

$$\hat{R}_t : \mathfrak{X}^A \cup \mathfrak{X}^B \rightarrow \{0, 1\}.$$

We assume that $\hat{R}_t$ provides sufficient information about R to allow an individual buyer or the seller to reach a best prediction for how R would classify circumstances that R has not yet publicly classified. $\hat{R}_t$ could consist of a series of past decisions in concrete circumstances or a set of incomplete rules for classification, for example, together with a method for reaching the best prediction of how past decisions or incomplete rules will generate classifications in novel circumstances. We assume that $\hat{R}_t$ enables both buyers and the seller to reach a unique common knowledge classification of any performance vector, but this expected classification may diverge from the one that will actually be announced publicly by the classification institution in a future period.

2.2.3. Sequence of play

The sequence of play for S and buyer A in each period t is as follows (the sequence for S and B is the same and happens simultaneously):

- A decides whether to purchase and pay P.
- If A purchases, S learns if it has the option of taking a cost-cutting action; the probability that an option arises to take a cost-cutting action that would also be judged wrongful by A is θ^A.
- S chooses a performance, X_t^A, deciding whether to engage in cost-cutting or not if the opportunity has been realized. S can observe $R(X_t^A)$ privately prior to acting.
- A and B observe X_t^A.
- A announces "X_t^A is wrongful" if $I^A(X_t^A) = 0$ and nothing otherwise.
- If A announces "X_t^A is wrongful" the classification institution announces "X_t^A is wrongful" if $R(X_t^A) = 0$ and "X_t^A is not wrongful" if $R(X_t^A) = 1$.

Observe that the classification institution is only asked to publicly announce a classification of a performance if a buyer judges the performance to be wrongful on the basis of its idiosyncratic logic.[9]

Using $\hat{R}_t$ buyer i can assess the frequency with which R classifies as wrongful a performance that I^i classifies as wrongful. Let r_t^i be buyer i's estimate of this frequency, which we will call the expected *convergence* between i's idiosyncratic logic and the common logic R. As a subjective probability, r_t^i is private information for i. r_t^i is updated each period as a new presentation of R is made public.

2.2.4. Equilibrium strategies and beliefs

We want to show that if R possesses certain characteristics, then it is capable of supporting an equilibrium in which the buyers coordinate on boycotting the seller when the seller engages in performances that are wrongful according to R. (We call these performances R-wrongful.) In this equilibrium, both buyers are better off than if they failed to coordinate boycotts.

First, as our reference point, note that the expected payoff starting in a period τ over the remaining (infinite) horizon of the repeated game for buyer i if the buyers never coordinate on boycotts is

$$\sum_{t=\tau}^{t=\infty} \delta^{t-\tau}((1-\theta^i)V - P). \tag{5}$$

9 The buyer has no incentive to seek a public classification that deems conduct wrongful and that the buyer judges to be not wrongful: the buyer receives V if the performance is not wrongful and cannot do better than this in our setup. We thus abstract from any strategic incentive that might arise in a system in which a buyer is awarded damages if a performance is judged wrongful.

Now suppose that buyer i believes that R-wrongful performances will prompt a two-buyer one-period boycott, and that this is common knowledge. Given the assumption (4) that the seller is deterred by this threat from taking the opportunity to cut costs by engaging in an R-wrongful performance, buyer i therefore expects the following payoff beginning in period τ:

$$\sum_{t=\tau}^{t=\infty} \delta^{t-\tau}((1-(1-r_\tau^i)\theta^i)V - P). \tag{6}$$

It is straightforward to see that (6) is weakly higher than (5). Note also that the extent to which (6) exceeds (5) depends on how close r_τ^i is to 1, that is, how often, given that the seller faces an opprtunity to engage in cost-cutting vis-à-vis i, R classifies as wrongful a cost-cutting performance that I^i also classifies as wrongful.

Now consider the following strategy, call it strategy R, for each buyer:

Purchase good in period t if and only if either
$R(X_{t-1}^i) = 1, i = A, B,$ or
no purchases were made at $t-1$.

Strategy R calls for each buyer to engage in a one-period boycott of the seller if an R-wrongful performance was observed (involving either buyer) in the previous period. Given assumption (4), a seller who anticipates that both buyers will follow this strategy will avoid R-wrongful performances.

We now want to consider a buyer's incentive to pursue strategy R. Suppose that buyer i believes that by engaging in a boycott in period τ the buyer can secure a switch in all future periods from a no-deterrence equilibrium to an equilibrium in which the seller is deterred from all R-wrongful performances. Boycotting under those beliefs is optimal for buyer i if

$$\begin{aligned}\sum_{t=\tau+1}^{t=\infty} \delta^{t-\tau}((1-(1-r_\tau^i)\theta^i)V - P) \\ > (1-\theta^i)V - P + \sum_{t=\tau+1}^{t=\infty} \delta^{t-\tau}((1-\theta^i)V - P).\end{aligned} \tag{7}$$

Equation (7) will be satisfied so long as

$$r_\tau^i > \frac{1-\delta}{\delta}\left[\frac{(1-\theta^i)V - P}{\theta^i V}\right]. \tag{8}$$

Inequality (8) shows that the more likely it is that the buyer will suffer a wrongful performance when the seller is undeterred (the higher is θ^i), the less good R

must be, from the buyer's perspective, at identifying wrongful performances. Similarly, the greater the loss associated with wrongful performance (the higher is P) or the more patient the buyer is (the higher is δ), the lower the convergence required between R and I.

We will say that R is *sufficiently convergent* for buyer i if buyer i has an incentive to engage in a one-period boycott given the buyer's beliefs about the extent to which boycotts will deter R-wrongful performances in the future. That is, for the cost of the boycott to be justified, R must sufficiently often classify as wrongful the performances that buyer i, using its idiosyncratic logic, judges to be wrongful. Inequality (8) establishes the criterion for sufficient convergence when buyer i believes that a boycott will fully deter all R-wrongful performances in all future periods.

Now suppose both buyers were known to be playing strategy R. The above analysis tells us that by both playing strategy R the buyers can secure a Nash equilibrium in a repeated game if r^i satisfies (8) for both of them. Unlike the proposed Nash equilibrium supported by strategy I, strategy R is implementable in the sense that both buyers and the seller can determine whether a particular performance does or does not trigger a boycott. R supplies a common definition of "cheating", although it does not necessarily align with either buyer's idiosyncratic logic. But does strategy R support a subgame perfect Nash equilibrium, in which it would be rational to boycott if a wrongful performance was observed (out of equilibrium)? We argue that it does, and precisely because the buyers hold idiosyncratic–and hence private–assessments of what counts as wrongful "cheating" vis-à-vis their own contracts.

Recall that r_t^i is reevaluated in each period, as the buyer observes a new public announcement about R, $\hat{R}_t$. This raises the possibility that even if it is rational for the buyer to boycott in period τ to secure a switch from a future with wrongful performances to one in which R-wrongful performances are deterred, it may not be rational at some future date, $t > \tau$. Because r_t^i is private information for buyer i, even if the buyers arrived at a state in which each expects the other to play strategy R, they cannot know that the other will play strategy R forever. R may be sufficiently convergent for A today, but will it be tomorrow? This incomplete information–about whether strategy R continues to be an optimal strategy for a buyer–generates an incentive for the buyers to carry through on boycotts (out of equilibrium or, more realistically, if the seller "trembles" [Selten 1975] and makes a mistake.)

Formally, we propose equilibrium beliefs as follows:

B1 *All agents believe that a buyer will boycott an R-wrongful performance in period t if and only if the buyer evaluates R to be sufficiently convergent in period t.*

B2 *Buyer j and the seller believe that R is sufficiently convergent for buyer i with probability* $(1-\rho^i)$, $\rho^i > 0$, *in period* 1 *and in period* $t > 1$ *if buyer i has played strategy R in all periods* $\tau < t$ *and with probability otherwise* 0.

Suppose that under B2, buyer *A*, for example, believes that with probability ρ^B buyer *B* will not boycott an *R*-wrongful performance. Under B1, if this event is realized and buyer *B* does not boycott, then this will cause both buyer *A* and the seller to infer the *R* is not sufficiently convergent for *B*. *A* and the seller will then update their estimate of the likelihood that *B* will boycott in the future to be 0. (Recall that each buyer estimates in each period, based on the current announcement $\hat{R}_t$, the expected convergence over all possible performances. A determination that *R* is not sufficiently convergent to warrant boycotting today implies a determination that *R* is not expected to be sufficiently convergent to warrant boycotting in any future period.[10]) Holding that belief, the seller now clearly will not be deterred from engaging in an *R*-wrongful performance the next time the opportunity arises. This generates an incentive for *B* to boycott: by boycotting, *B* can preserve B2, that is, the belief by buyer *A* and the seller that with probability $(1-\rho)^B$, *R* is sufficiently convergent for *B*. Conversely, a decision not to boycott is a decision to ensure that the other buyer and the seller do not expect any effective boycotts to arise in the future. Such a future is one in which there is no deterrence.

B1 and B2 modify slightly the detemination of what it takes for *R* to be sufficiently convergent from the derivation reached in (8) above. A buyer *i* who entertains beliefs B1 and B2 will find it optimal to boycott in period τ if

$$(1-\rho^j)\sum_{t=\tau+1}^{t=\infty}\delta^{t-\tau}((1-(1-r_\tau^i)\theta^i)V-P)+\rho^j\sum_{t=\tau+1}^{t=\infty}\delta^{t-\tau}((1-\theta^i)V-P) > (1-\theta^i)V-P+\sum_{t=\tau+1}^{t=\infty}\delta^{t-\tau}((1-\theta^i)V-P). \tag{9}$$

Inequality (9) is satisfied if

$$r_\tau^i > \frac{1-\delta}{\delta}\frac{1}{1-\rho^j}\left[\frac{(1-\theta^i)V-P}{\theta^i V}\right] \equiv \underline{r}^i. \tag{10}$$

Inequality (10) imposes the additional requirement, relative to (8), that for buyer *i* to face an incentive to boycott, ρ^j cannot be too high. That is, each buyer must believe that the likelihood that *R* is sufficiently convergent for the other is not too low. $\underline{r}^i$ establishes a lower bound on the degree of convergence

10 The agents in this model do not treat *R* as dynamic: they do not have predictions about how *R* might change in the future. Their estimates of convergence can change over time not because *R* changes but because their information about *R* changes.

that a buyer must anticipate in order for the buyer to be willing to participate in boycotts of wrongful performances.

Introducing the possibility that R may fail to be sufficiently convergent to generate an incentive for a buyer to boycott also modifies the seller's incentive to avoid R-wrongful performances from our initial derivation (4) where a boycott response was certain. Suppose the seller entertains the same beliefs as buyer i about the likelihood that R is sufficiently convergent for buyer j. Then, the seller expects that an R-wrongful performance will prompt a two-buyer boycott in the next period with probability $(1-\rho^i)(1-\rho^j)$. The seller will then be deterred from R-wrongful performances only if

$$c < \frac{2\delta(1-\rho^i)(1-\rho^j)}{1+2\delta(1-\rho^i)(1-\rho^j)}P. \tag{11}$$

Inquality (11) is a more relaxed constraint than (4) as we would expect: the expected penalty for an R-wrongful performance is strictly lower when $\rho > 0$ and so the seller is willing to incur the penalty for lower values of c, the payoff the seller enjoys from engaging in wrongful performance.

We can now state our result:

Proposition

If R *is sufficiently convergent for both buyers and*

$$c < \frac{2\delta(1-\rho^i)(1-\rho^j)}{1+2\delta(1-\rho^i)(1-\rho^j)}P$$

then the following strategies and beliefs support a perfect Bayesian Nash equilibrium in which both buyers boycott R*-wrongful performances and the seller does not deliver* R*-wrongful performances:*

Buyers' strategy: *Play strategy* R *in any period* t *unless the other buyer has failed to play strategy* R *in some period* $\tau <$ t.

Seller's strategy: *Restrict performances to the set* $\{X_t^i \ni R(X^i) = 1\ \forall i, \forall t\}$ *unless a buyer has failed to play strategy* R *in some period* $\tau <$ t.

Beliefs (all players): *(B1) Buyer* j *will boycott an R-wrongful performance in period* t *if and only if* R *is evaluated by* j *to be sufficiently convergent in period* t*, that is, if*

$$r_t^j > \underline{r}^j.$$

(B2) R *is sufficiently convergent for buyer* j *in period* t *with probability*

$$= \begin{cases} (1-\rho^j),\ \rho^j > 0 & t=1 \text{ and } t>1 \quad \text{if buyer } j \text{ has played strategy } R \quad \forall \tau < t \\ 0 & \text{otherwise.} \end{cases}$$

For these strategies and beliefs to support a perfect Bayesian Nash equilibrium—in which the threat to boycott is credible off-equilibrium—we need to show that the strategies of all players are sequentially rational and that beliefs are consistent with the equilibrium strategies. Most of the proof of this proposition follows directly from our analysis above showing that a buyer who entertains beliefs $B1$ and $B2$ is better off boycotting than not so long as R is sufficiently convergent as defined in (10). Failure to boycott in the off-equilibrium event that an R-wrongful performance occurs prompts both the seller and the other buyer to update $B2$ to a belief that R is definitely not sufficiently convergent for the non-boycotting buyer. Given this updated belief, it is optimal, as strategy R prescribes, for the other buyer not to boycott in the future. This shows that the proposed equilibrium strategies are sequentially rational. It remains just to check that the proposed beliefs are consistent with the equilibrium strategies. This is straightforward: in equilibrium a buyer will boycott if and only if R is sufficiently convergent and so it is consistent for the other buyer and the seller to infer from a failure to boycott that the buyer has reached a judgment that R is not sufficiently convergent.

3. DISCUSSION

We have shown that a common logic R can support an equilibrium in which wrongful conduct that destroys value is effectively deterred by decentralized collective punishment, that is, in the absence of a centralized coercive body. The common logic—an institution that implements a system for classifying actions as wrongful or not—achieves this by doing two things. First, it coordinates expectations about how performances will be classified. Second, it supports a buyer's incentive to participate in boycotts of performances that the logic deems wrongful, even when those are wrongs suffered by the other buyer.

Our claim is that the equilibrium coordinated by R can be usefully interpreted as a *legal order*, despite the absence of a centralized coercive force. As a starting point, we suggest that for an equilibrium to be characterized as a legal order it should, at a minimum, have the following characteristics, all of which are displayed by the equilibrium coordinated by R. First, the equilibrium must display order: behavior systematically follows a designated pattern. In our equilibrium, we observe this: buyers enter into contracts with sellers, pay them up front and receive deliveries of goods. Second, the patterning of behavior must be based on a normative classification. By this we mean that there is an evaluative framework that values the elicited behavior more highly than alternative behavior. In the equilibrium organized by R, the buyers collectively prefer the behavior elicited by the system—punishment and hence avoidance

of *R*-wrongful performances—to the alternative. Third, the content of the classification system must be capable of being deliberately articulated by a third-party institution.

These criteria for a legal order identify legal order as a form of deliberate, as opposed to spontaneous or emergent, order. This distinguishes legal order from, for example, the emergence-in-fact of an equilibrium set of property norms such as those analyzed by Sugden (1996). Order in these spontaneous cases is the result of repeat interaction and the confluence of individual decisionmaking exercised in the absence of external coordination; any normative classification is limited to the classification supplied by individuals acting independently.

We suggest that a legal order requires the capacity for an individual or entity to announce a classification system. The capacity to deliberately articulate, and hence clarify or change, the content of *R* satisfies H.L.A. Hart's concept of law, which he defines as the presence of secondary rules that determine the validity of primary rules (Hart 1961/1997). The capacity to articulate, clarify, and adapt the content of a classification system is fundamental to a concept of law that serves the needs of economists and political theorists, we suggest, because these social scientists are interested in the distinctive role of law as a vehicle for policy and politics.

We do not claim that all deliberate social orders are legal orders. A tyrant can establish a deliberate social order, for example, and most such orders are not usefully called legal orders. Our claim is that a legal order is a species of deliberate social order, and in particular a deliberate social order that possesses several attributes that are commonly associated with the existence of law or the rule of law. Fuller (1964), for example, defines law as "the enterprise of subjecting human conduct to the governance of rules." His concept of law as an "enterprise" is consistent with our notion that the classification on which legal order is based must be one that is capable of being deliberately chosen and changed. Fuller then goes on to identify eight characteristics that rules must possess in order to be properly called "legal" rules. These are: generality, stability, prospectivity, promulgation, clarity, non-contradiction, congruence (between rules as announced and rules as applied), and possibility (the rules do not call for actions that are not feasible for the subjects of the rules). Other legal philosophers (e.g., Raz 1977) see these characteristics as definitive of, if not legal order per se, the existence of the rule of law. Still others urge that the list of legal attributes should be expanded to include process characteristics such as the availability of open courts that engage in formal argument, processes, and practices to apply public-oriented rules (Waldron 2008). The equilibrium order secured by *R* in our model possesses several of the features these philosophers emphasize as definitive of "law".

3.1. Generality, Stability, Prospectivity, and Congruence

A rule is general if it is articulated in terms of categories or principles that can be applied to specific factual circumstances not fully described by the rule itself. A statement of the form "all deliveries made more than 4 days after the contract date shall be deemed to be late" is general: we do not need all the details about who the buyer and seller are, what is being sold, other terms in their agreement, the date of their delivery, and so on to classify a delivery as wrongful or honorable. A classification scheme might be general without containing articulate general rules of this form, however. A collection of classifications reached in the past, for example, will be general if those classifications are generalizable: if by studying those classifications and analyzing the reasons for them and the decisionmaking rules of the classification entity a person can make a reasonable prediction about how a comparable but as-yet-unobserved performance vector would be classified. This is how the common law works, for example. According to traditional Anglo-American legal concepts, the common law is found and not made: its rules and principles are *immanent*, it contains all of its principles even if they are not articulated until a specific case is adjudicated.[11] Decisions that cannot be reached until after specific circumstances arise are considered to be law-like and not *ad hoc* because of this understanding of what it means for the law to be generalizable in a principled way.

In our model, a system of classification lacks generality if it is not possible to predict with sufficient accuracy the classification of a significant number of performance vectors without obtaining a specific declaration from the classificatory entity, R. The generality or generalizability of the common logic R is captured in our model by the idea that buyers can use the public announcements about R, namely $\hat{R}_t$, to predict how R will classify performances in the future in circumstances that may be unanticipated in detail by anyone other than the buyer. An R that satisfies the criterion of generality is capable of generating relatively high values of expected convergence, as measured by r, for each buyer, as required in our equilibrium. If R were not general, public information about R would not provide a basis for the buyers to reach relatively high levels of confidence about the likely convergence between their idiosyncratic logic and the common logic. The capacity to generalize from publicly announced features of R is critical in our model in light of our assumption that buyers possess private information about the circumstances they are likely to encounter. This rules out the possibility that R is an omniscient classification

11 Blackstone held that it was not the judicial function to "pronounce a new law, but to maintain and expound the old one". 1 William Blackstone, *Commentaries* 69.

scheme, generating an ex ante listing of the classification of all possible performances in all possible circumstances.

The relationship between generality and prediction of future classifications brings into focus other characteristics of R that are usually associated with the concept of legal order. In order for buyers to conclude that R is sufficiently convergent with their idiosyncratic logic, as required in equilibrium, it must be the case that R is relatively stable in the sense that $\hat{R}_t$ is believed to be capable of supporting sufficiently reliable predictions about R in the future. If R were unstable—if it were known that R is subject to wide swings in content—then current announcements about R would not support expectations of sufficient convergence. Similarly, R must be reasonably prospective in the sense that the classifications R announces today must be expected to apply to future conduct. This is necessary both for buyers to consult their estimates of R to evaluate their payoffs in future periods, and for the seller to make compliance decisions as required in equilibrium, calculated to avoid boycotts by avoiding R-wrongful choices. An R that systematically announced unpredictable classifications for past conduct could neither guide the seller's performance decisions nor structure buyer expectations and strategies so as to support equilibrium deterrence of R-wrongful conduct. And, relatedly, if R does not display reasonable congruence between rules and principles as announced and as applied—if $\hat{R}_t$ is not a good predictor for R—equilibrium deterrence of R-wrongful conduct cannot be supported.

3.2. Universality

Scholars writing in contexts related to law use the term "universality" in a number of different ways. Some, following Kant, use the term to describe an obligation that, if it applies to person A in circumstances X, applies to all persons in circumstances X. A legal rule might be described as "universal" if it applies to all persons within some jurisdiction, although the rule may itself designate specific attributes that a person must have in order to be covered by the rule. (A rule for vendors who supply a particular product or are located in a particular location or attain a certain market share, for example.) This concept of universality thus overlaps with ideas of equal treatment of persons who are in relevantly similar circumstances and of the impersonal application of rules, that is, without regard to (irrelevant) personal characteristics.

We use the term universal in this article in a different way, distinguishing it from concepts of equal treatment and impersonal application of rules. First, we use the term to describe an attribute of a system of logic, rather than particular rules in that system. Then, borrowing from the idea of a universal tool or procedure that can help to resolve all problems in a particular class, we say that

a logic is universal if it addresses (one way or another) the rights and obligations of all individuals. A universal logic, then, is like a book of rules that includes entries for all the possible configurations of persons and relationships. A system of universal logic need not contain rules that call for equal treatment of persons or that treat personal characteristics as irrelevant. The rules in a universal logic that apply to the circumstances of person A need not be the same as those who apply to person B, even if A and B are in similarly situated. (We do not mean to say that it would not be more likely as an empirical matter for a universal logic to treat A and B the same; just that this is not formally implied by our definition of universality.)

In our model, the logic R must be universal in the following particular and qualified sense: both buyers can identify rules in the logic that address their particular interests and the circumstances they anticipate in their relationships with the seller. Mathematically, this means that R must be defined over the full set of circumstances that either A or B considers relevant: $R: \mathfrak{X}^A \cup \mathfrak{X}^B \rightarrow \{0, 1\}$.

Universality in our model derives from the equilibrium requirement that the common logic must be sufficiently convergent with both A and B's idiosyncratic logic to attract both to participate in a coordinated boycott triggered by the application of the common logic. If, for example, R only classifies actions when they involve A, B will refuse to participate in the boycott.

Whether a logic that is universal in our sense would also display equal treatment is then a function of how a particular equilibrium R emerges and remains stable, something we have not modeled here. We conjecture, however, that the chances of establishing R as a stable equilibrium will be greater if R is addressed to abstract persons or entities, rather than specific individuals. Relatedly, to the extent that even heterogeneous agents can face similar circumstances which they judge in similar ways (both buyers, for example, are likely to judge a complete failure to deliver any goods as wrongful), we conjecture that, with little ability to predict the particular content of each buyer's idiosyncratic logic, the classification institution can increase the likelihood that both buyers find R to be sufficiently convergent by designating a common logic that does not discriminate between A and B in its classification of some performances.

It is important to emphasize that our analysis of universality is not based on an assumption that agents prefer fair or equal treatment per se. Our buyers derive utility only from the transactions they engage in with the seller. They do not enjoy community benefits or good feelings about themselves or the goods associated with conformity to norms per se. Similar to Binmore's (1994, 1998) effort to ground the Rawlsian "justice as fairness" principles in game theory, our analysis grounds the emergence of "general" rules on the interaction of self-interested agents who do not possess an inherent set of values over their relative treatment by the rules.

It is also important to note that universality in our model serves to support the punishment incentives of agents whose participation in punishment is necessary for effective deterrence. In this sense, it is a qualified universality. If we were to add a third buyer *C*, and continue to suppose that effective deterrence still only required a two-buyer boycott by *A* and *B*, then there would be no need for *R* to cover *C*'s interests. So if *A* and *B* are members of the elite, for example, and *C* is a peasant, nothing in our result would rule out a "universal" common logic that deemed performances wrongful only if the injured buyer is a member of the elite. Conversely, if effective punishment requires *C* to boycott as well (if, for example, there is an equal probability that the seller will only have an opportunity to sell to any two of the three buyers in the next period), then an equilibrium *R* will be universal with respect to *C* as well.

3.3. Clarity, Non-contradiction, Uniqueness: Authoritative Stewardship

The model assumes that it is common knowledge that, once *R* has been established in equilibrium, all agents agree on what constitutes an *R*-wrongful performance and hence all reach the same prediction about the likelihood of an effective boycott in response to a given performance in the future. *R* is thus assumed to produce unambiguous classifications—which requires that *R* produce classifications that are both clear and unique. Implicitly, it also requires that *R* itself be unique, that is, that all agents are consulting the same common logic to assess wrongfulness.

Clarity and uniqueness impose constraints on the structure of the reasoning employed by the logic when accurately applied: There must be, at least in theory, a "right" answer to the question of whether a particular performance is wrongful or not. The logic must be coherent and not contradictory. Unique classification does not imply that the rules and principles that make up the logic produce an *obvious* classification. The set of rules and principles that comprise the logic could be complex and ambiguous and capable of producing multiple answers, although this would make it more costly. (Our simple model assumes all logics are costless to use.) Agents may make errors in applying the logic. What is important is that there be a recognized process for determining a unique answer among a set of possible answers implied by the rules and principles. This observation gives content to our original definition of a logic not merely as a set of rules or principles but rather as the product of a third-party institution. Achieving a unique common knowledge classification necessitates that there be an *authoritative steward* of the classifications reached by the logic: a unique arbiter able to resolve complexities, ambiguities, and gaps. This sheds light on why we generally find that an established legal

system in a complex environment usually has a single Supreme Court, for example.

In our model, we have presumed that a definitive classification can be obtained from the authoritative steward—the classification institution—before the seller or buyers have to choose their actions. Thus we can only claim to have shown that clarity and uniqueness are sufficient to support an equilibrium under R with effective deterrence. We conjecture that an equilibrium can also be supported even if we allow some degree of noise or ambiguity in classification. But our model provides insight into the likely limits on the extent of ambiguity that can be supported. Consider what happens in the event that the agents reach different classifications of a particular performance. Suppose in particular that the seller classifies as not R-wrongful a performance that the buyers classify as R-wrongful. The proof of our Proposition takes care of this case: because the equilibrium is perfect, we know that although the equilibrium calls for the seller never to make an R-wrongful delivery, the buyers will nonetheless respond to the R-wrongful delivery by carrying through with an effective boycott. Moreover, we know from the set up of the model that each buyer is willing to participate in the equilibrium despite the risk that some performances the buyer judges to be wrongful will occur; this is what we capture with the concept of sufficient convergence. It is straightforward to see that we can reinterpret our measure of expected convergence, r_t^j, to take into account the risk that even if R classifies a performance as wrongful, there is a chance that it will not be deterred. The intuition of sufficient convergence gives a basis for conjecturing that, so long as this risk is not too great, then buyers will be willing to forego profits in some periods in order to protect the future benefits of deterring a sufficient number of wrongful deliveries. Conversely, as ambiguity increases and the risk that the seller and buyers reach different predictions about the wrongfulness of particular conduct, the estimates of r by the buyers will fall and eventually equilibrium cannot be sustained.

The more subtle case involves the risk of different classification of performances by the buyers. Here, the problem is not merely the introduction of a risk in any period that coordination will fail and an effective boycott will not result—the model includes the potential for such risk (ρ) arising from the recognized potential for a buyer to update its expectations about the convergence between R and I^j based on new public announcements $\hat{R}$. The more difficult problem generated by ambiguity in classification among the buyers is the impact of ambiguity on the interpretation of a failure to boycott. In the case that a unique common knowledge classification exists, only one consistent inference can be drawn from buyer j's failure to boycott: R is not, or is no longer, sufficiently convergent to support j's participation in coordinated boycotting. If it is common knowledge that in applying R the buyers, in some

cases, will reach different classifications of performance, then this inference is not warranted. Our model suggests that an equilibrium could be sustained in which the buyers do not update their beliefs about the likliehood that *R* is no longer sufficiently convergent for a buyer until they have observed a number of failed boycotts, but we have not shown that. But our intuition suggests that there will be, potentially sharp, limits to the extent to which coordination can be sustained in the presence of ambiguity. Moreover, it seems safe to conjecture that, given that ambiguity will trigger mistakes in boycotting and require more periods of costly boycotting to convey information about the extent to which *R* is or remains to be sufficiently convergent for a buyer, we will be more likely to see the emergence and stability of common logics that more effectively reduce ambiguity through institutional attributes that secure clear and unique classifications.

Although we have not modeled a process by which *R* is developed or proposed, it seems clear that the probability that *R* succeeds in establishing a deterrence equilibrium is higher the more effectively *R* reduces ambiguity. Similarly, if institutions are competing for selection, the stewards controlling a common logic will be more likely to secure selection of their institution if they more effectively achieve unique and clear classifications.

3.4. Impersonal, Neutral, and Independent Reasoning

Our model has only two buyers and one seller. It may therefore seem reasonable to suppose that an equilibrium could be supported even if *R* classifies performances on the basis of an idiosyncratic logic, using reasoning that cannot be reproduced by the buyers and the seller. The model would only require that the buyer and seller be able to query the institutional agent to learn the classification for a particular performance. But a query-based system in practice is likely to be costly; it will also involve disclosing to the institution private information that a buyer may prefer to keep private unless and until there is a need for a public classification. Moreover, in a more general model with a large number of buyers and sellers, the capacity for a single agent to respond to queries is likely ultimately to be exhausted.

We therefore interpret the model to suggest the importance of a logic based on *impersonal reasoning*. By impersonal reasoning, we mean that the operation of the logic on a set of facts regarding a performance produces a classification that is invariant to the identity of the person or entity engaged in the operation.[12] Impersonal reasoning implies that the institution providing the logic

12 This does not necessarily mean that agents do not differ in their competence in employing the logic: although we have assumed that classification is costless, a more general model could sustain some

must be *neutral and independent*: the agents who provide the classifications of R must have no interest in those classifications. This suggests a strong reason to believe that one of the buyers cannot just propose using its own idiosyncratic logic as the basis for boycotting.

Neutrality and independence are routinely identified by legal philosophers as key attributes of systems that observe the rule of law. These accounts often ground the requirement of neutrality in a normative principle such as fairness. Raz (1977, 201) offers an informal behavioral reason for neutrality: "it is futile to guide one's action on the basis of the law if. . . the courts will not apply the law and will act for some other reasons". Our model suggests a different reason for neutrality: neutrality reduces ambiguity. A lack of neutrality undermines the capacity of the law's classification system to coordinate effective deterrence by increasing the cost of and/or variance in classification.

3.5. Public Reasoning and Open Process

In developing the model, we assumed that the logic R is *publicly accessible*: both the buyers and the seller have access to the public announcements $\hat{R}$ and specific declarations about the application of the logic when making their decisions about boycotting and performance. More subtly, however, the model implies that the publicness of the logic goes beyond mere publication of the rules, as most legal theory presumes.[13] Our model suggests that a robust common logic is likely to come in the form of *public reasoning* elaborated in an *open process* to which an interested party may introduce their private information and reasoning. This accords with Waldron (2008), for example, who insists on the importance of argument in the concept of law; argument presumes the capacity for an interested party to bring their own understanding of the case to the attention of the court and to have that potentially incorporated into the authoritative reasoning of the court.

Both public reasoning and open process in our model find their root in the heterogeneity and idiosyncrasy that generates the problem of ambiguity and the need for a common logic in the first place. Put differently, agents in a homogeneous world with shared and unambiguous classifications of all performances as "cheating" or not, have no need for an external institution to provide a common logic; in such a world, we can predict, as do

costs to hire the services of an expert interpreter of the logic (such as a lawyer). But the logic would still have to consist of impersonal reasoning in the sense that the classification reached by an expert did not depend on the identity of the expert.

13 "The law must be open and adequately publicised. If it is to guide people, they must be able to find out what it is." Raz (1977, 198–199)

Hume (1739–40/1978) and Sugden (1996), that norms to coordinate behavior will spontaneously emerge. Milgrom, North, & Weingast (1990) find that all that is needed in such a world is an institution that serves to share information across traders separated in time.

The likelihood that a common logic *R* will be characterized by open and public reasoning follows from our model because the assumption of idiosyncrasy suggests that the classifications reached by the logic must be general, as we discussed earlier. Recall that we have defined each buyer's idiosyncratic logic as an inaccessible reasoning process that maps (potentially private) information into an assessment of the value of a potentially complex set obligations on the seller.[14] Having no access to the idiosyncratic reasoning of individual buyers when it offers its logic as a candidate coordination device, a third-party institution must provide a logic that is capable of integrating, coherently, the information and reasoning from individual buyers through an infinite horizon. The logic, therefore, cannot be (just) a data set collecting classifications already reached by the logic; it requires placeholders for dealing with as-yet-unimagined circumstances. Nor can it be a complete prescription of how all possible circumstances would be classified by the logic. To do we this would require access to the idiosyncratic logic of (possibly as-yet-unknown) buyers who are uniquely able to assess the value and intended content of their transactions with sellers. As we have seen, the logic *R* must be sufficiently convergent with each buyer's idiosyncratic (ex ante inaccessible) logic in order to attract the buyer's participation in the coordination equilibrium.

In a system in which classifications are based on general principles and/or generalizable decisions, classification requires elaboration in particular circumstances. In our model, those particular circumstances are initially private information. To elect to participate in the boycott equilibrium, each buyer must be able to elaborate the logic privately as it applies to these privately known circumstances and considerations. We have already discussed the requirement that this elaboration produces a unique classification. Ultimately, when this set of circumstances becomes relevant (the seller is contemplating a potentially wrongful performance or the buyers are determining whether to engage in a boycott in response to a potentially wrongful performance), this classification must be capable of becoming public.

Thus we conjecture that in a stable classification system, the elaboration of the reasoning—its application to particular circumstances—will be conducted in public and in a manner open to a presentation from the initially privately

14 In a world where delivery within four days of a contract date is generally considered acceptable, for example, buyer *B* may be an innovative manufacturer that has discovered how to employ just-in-time delivery or variations in wholesale packaging to improve the allocation of inventory.

informed buyer (more generally, also the seller) of how its idiosyncratic reasoning plays out in the common logic. In our model, buyer *A* does not care about buyer *B*'s idiosyncracies unless and until *B* is the potential victim of a wrongful performance and *A* has to decide whether to boycott or not. At that point, *R* is presumed to include an open and public reasoning process to determine, uniquely, whether the performance is *R*-wrongful or not. More generally, we predict that a classification institution that is open to hearing from individual buyers and sufficiently public is likely to give buyers greater confidence that *R* will, in practice, converge sufficiently with their idiosyncratic logics.

4. RELATIONSHIP TO THE LITERATURE

4.1. Philosophy of Law

Although we have appealed to some legal philosophical concepts in our discussion above, we do not intend our work to be a philosophical contribution to the extensive literature in analytical jurisprudence that has considered the question in depth of "what is law". The participants in that literature frame their work in terms of the relationship between law and morality, often from the internal perspective of an agent within a legal system. We are not engaged in moral theory, or even normative theorizing, in this article. But, as Kornhauser (2004) has noted, some of those clearly recognized as major contributors in analytical jurisprudence—such as H.L.A. Hart and Lon Fuller—can also be seen as progenitors of the project we take up, of developing a social-scientific concept of law. It is therefore important to sketch out how we think our work relates to legal philosophy.

Modern positivists distinguish between the concept of law per se and the concept of the rule of law. In its sharpest formulation, this distinction emphasizes that the concept of law is devoid of any necessary normative content; it is an effort to capture what, in fact, constitutes "law" regardless of whether the content of a legal system is judged to be good or bad. In contrast, the rule of law is a normative ideal: a legal system may or may not display the desirable qualities of the rule of law. Fuller (1964), for example, argues that to be recognizable as law, legal rules must be characterized (more or less) by the eight characteristics we listed earlier: generality, promulgation, prospectivity, clarity, non-contradiction, feasibility, stability, and congruence between rules as announced and rules as applied. Raz (1977), on the other hand, argues that beyond some minimum these features are not necessary to the existence of law per se but rather are virtues displayed by the rule of law.

The distinction between the rule of law and the concept of law takes on special importance for legal philosophers who are engaged in the project of determining the relationship between law and morality and in particular the relationship between the existence of a legal rule and the reasons for action that law gives a person to whom the rule is addressed. This is a largely internal point of view. From this point of view, answering the question of "what is law" is a matter of determining what counts as a valid law for purposes of those within a legal system who seek to be guided by the law—judges, officials, and ordinary citizens. If an unjust law is not a law then it does not give rise to a legal obligation for those who seek to be guided by the law. If a valid law is determined by a system of social validation that depends, for example, exclusively on compliance with particular procedures and not on the substantive content of the law, then whatever reasons law gives for complying with its rules are independent of whatever moral reasons we might have for complying with a rule or acting in any other way.

We do not emphasize the distinction between the concept of law and the rule of law in this article because we adopt an external perspective on law: how are we to understand the phenomenon of law as a mode of social organization? What criteria for distinguishing legal order from other types of order will aid in the effort to predict and identify the emergence or disappearance of distinctively legal order? Most centrally, what are the mechanisms by which law achieves order, how can these mechanisms be structured or modified to achieve particular ends and can these mechanisms sustain legal order as an equilibrium? The definition of law that we work toward is to be judged by its success in helping to frame a theory of law as a form of social organization distinct from other forms of social order.

As Kornhauser (2004) notes, this social-scientific approach to developing the concept of law shares important common ground with the positivist account in analytical jurisprudence. Both Fuller (1964) and Raz (1977) ground several of their arguments for why law, or the rule of law, must possess certain characteristics on an informal model of human behavior. Both presume, for example, that, as a practical matter, people cannot plan on the basis of rules that they cannot discover or ones that they do not expect to govern the application of future penalties and therefore conclude that legal rules must be stable, publicized, and largely prospective. Hart (1961/1997) emphasizes that what counts as valid law in a given community is ultimately a matter of social fact that cannot be determined through moral reasoning or semantic analysis but only through the decidely non-normative analysis of social interaction in practice. Moreover, in what Kornhauser (2004) calls an "abandoned project of descriptive sociology", Hart motivates his concept of law—which he identifies by the presence of a set of secondary rules that determine the validity and modes of

application of primary rules—with an appeal to the challenges that face a society that is under pressure to adapt its primary rules to changes in the environment or increases in complexity or heterogeneity.

Our approach can be seen as an effort to pick up this starting point for a social-scientific theory of the phenomenon of legal order. We share with Hart the intuition that the emergence of legal order is linked to increasing complexity and heterogeneity in human environments and the pressure this puts on spontaneous social order. Our contribution is to take this insight more firmly in the direction of social scientific, particularly rational choice, analysis.

4.2. Coordination Accounts of Law

A large literature in both social science and legal philosophy, going back to Hume (1739–40/1978), explores the idea that law plays a role in coordinating behavior.

In legal philosophy, coordination accounts have been largely spurred by Hart's (1961/1997) claim that the validity of law is ultimately a matter of social convention: a rule counts as a legal rule if the participants in a given legal community believe and behave as if it were a legal rule. Lewis (1969) although not specifically focused on law, provides a key definition of convention: a regularity of behavior in which an agent perceives him or herself to be better off engaging in the behavior on the expectation that all others will do also. For Lewis, and the legal philosophers who followed him, a convention is a solution to a coordination problem in the sense of economist Schelling (1981). Postema (1982) argued that the practices of the officials in a legal system who, according to Hart's view, define what is valid law have the characteristics of a coordination problem and in this sense the secondary rules of a legal system can be understood as conventions that resolve this problem. Other philosophers examining the role of convention in understanding the validity, authority, and autonomy of law include Raz (1977); Finnis (1980, 1989); Gans (1981); Marmor (1998, 2009); and Green (1983). Although this literature, in places, appeals to formal game theory, it is largely focused on the relationship between a coordination account of law and the normativity of law in the sense of the capacity of law to generate moral reasons to obey the law.

Positive political theory and the law has long recognized the importance of coordination in one aspect of the law, namely, constitutional law with a focus on constitutional stability. Most new constitutions fail (Elkins, Ginsburgh, & Melton 2009), so why do those few survive? Hardin (1989, 2006), following Hume (1739–40/1978), argues that the central feature of constitutions is to provide coordination for citizens around various rules (see also Ordeshook 1992; Calvert & Johnson 1999). Constitutions, in this view, create focal

solutions that allow citizens to create order. In a model closely paralleling that in this paper, Weingast (1997) argues that constitutional stability requires that citizens have the ability to coordinate against governments that seek to transgress constitutional provisions. To do this, citizens must create focal solutions to the problem of what features of the constitution are worth defending. Constitutions that become focal points (typically in moments of crisis) have greater ability to survive then ones that do not. Similarly, Fearon (2011) argues for the coordination effect of elections in democratic (and hence democratic constitutional) stability.

In the economics literature, coordination accounts of law begin with coordination accounts of spontaneous social norms without deliberate design or legal institutions. Sugden (1996) uses focal point equilibria (Schelling 1960) to explain the spontaneous emergence of self-enforcing conventions about coordination, reciprocity and property rights to resolve rival claimants disputes. Binmore (1994, 1998) also approaches the problem of explaining the emergence of conceptions of justice—particularly fairness—as the resolution of a coordination problem in which the equilibrium must be self-enforcing. Dixit (2006) considers multiple settings in which coordination can be achieved by extra-legal conventions, including focal point settings.

Several authors extend the analysis of spontaneous social norms to law by arguing that where there are multiple self-enforcing coordination equilibria, law can serve as a focal institution to deliberately select an equilibrium (Cooter 1998; Basu 2000; McAdams 2000, 2005; Mailath, Morris & Postlewaite 2001, 2007; Myerson 2004). Like Sugden (and Hume), both McAdams and Myerson, for example, observe that a rule that deemed the immediate possessor of a piece of property to be its rightful owner can coordinate the strategies of rival claimants so as to avoid wasteful contests over the property. If both claimants expect the other to apply a concept of "rightful" ownership, then the "rightful" owner will rationally claim and the other will rationally recede. Whereas Sugden and Hume look to the spontaneous emergence of this rule, however, McAdams and Myerson consider the role for legal institutions such as a legislative assembly or adjudicator. Myerson (2004) proposes that an assembly can select generally understood principles to coordinate expectations about who will rightfully claim what. McAdams (2005) considers in depth the way in which adjudicators can convey information about facts or the prevalence of community beliefs about the content of a norm to support coordination on a particular equilibrium in the presence of ambiguity about a convention or its application. Myerson (2004) also considers the role for an arbitrator who recommends an equilibrium when general principles do not cover the situation or are ambiguous.

In all of these literatures—legal philosophy, positive political theory, and economic analysis of law and norms—the appeal to coordination is an

appeal to a very specific, and probably rare, payoff structure. This is the structure of a coordination game in which coordination is both necessary and sufficient to sustain a Nash equilibrium. The canonical examples used in the literature are Schelling's (1960) Meeting game (M), the Battle of the Sexes (BOS) game, and the Hawk-Dove (HD) game. In M and BOS, both agents enjoy higher payoffs when they choose the same strategy (go to Grand Central Station or go to the Empire State building; attend a play or attend a football game). In the Meeting game, the agents are indifferent about whether they go to Grand Central Station or the Empire State building, so long as they both go to the same place. In BOS, one agent prefers the equilibrium in which both agents attend a play and the other prefers the equilibrium in which they both attend the football game, but both prefer being together to being at different events. In HD, each agent would prefer to play Hawk (claiming a contested object) than to play Dove (conceding the contested object) but each also prefers the equilibrium in which he or she plays Dove and the other plays Hawk to one in which both play Hawk. In all three of these games, coordination of strategies is both necessary and sufficient for equilibrium. Sufficiency comes from the fact that the payoffs in these games are such that an uncoordinated strategy is never preferred to coordination. In this sense, the only role for a third-party institution is to achieve coordination. Once that is done, equilibrium is achieved.

In our model, in contrast, coordination is necessary for equilibrium, but not sufficient. Equilibrium requires more: specifically, equilibrium requires legal attributes that render a coordination equilibrium preferable for all agents to the payoffs that can be achieved without coordination. Put differently, our model does not presume the structure of a classic coordination game of the type that this existing literature assumes. This makes our model far more general as an account of the role of coordination in explaining legal order than anything offered in the existing literature.

A second distinction between our approach and the existing literature is that, with the exception of Basu (2000) and McAdams (2005), the existing coordination accounts of law focus on the coordination problem facing agents engaged in primary behavior: choosing the side of the road on which to drive, whether to claim contested property, or whether to apply a conventional interpretation of a statute, for example.[15] In our model, in contrast, the problem of coordination is one faced by agents who are potentially engaged in punishing primary behavior:

15 Basu (2000) emphasizes that official enforcers such as judges and police must also choose to comply with a given legal norm for the norm to establish an equilibrium. McAdams (2005) notes informally that law might also serve to coordinate punishment strategies to enforce legal rules. In both accounts, however, enforcers are presumed to be engaged in a coordination game in which coordination is necessary and sufficient for equilibrium.

responding to those who drive on the wrong side, take what is not theirs, or adopt unconventional statutory readings. That is, we focus on how the characteristics of legal rules governing primary behavior impact the coordination problem facing enforcers of those rules. This also makes our approach far more general than the existing accounts. As McAdams (2000) is careful to note, the expressive account of law (Sunstein 1996) is a partial account of law, applying only to those settings in which primary behavior happens to be characterized by an overriding incentive to coordinate; that is a setting in which no punishment is required to enforce compliance with a legal rule. In our account, we presume the far more ordinary setting in which a legal rule imposes a penalty on particular conduct and on that basis channels behavior in the direction of compliance.

Last, our approach introduces a level of formal modeling that is missing from the existing coordination accounts. Although this literature sometimes employs game theory, it does so by essentially making the claim that if interactions are structured as coordination games where agents are always better off coordinating than not, then law can provide a focal point to select among multiple coordination equilibria. The objective functions and information states of the agents in these games are not specified and there is no foundational account of when payoffs will be structured in this way. We build a formal model that derives the structure of payoffs based on foundational assumptions about utility and information. Our account therefore demonstrates, rather than presumes, the expected payoffs associated with different strategies.

4.3. Collective Punishment

Cultural anthropologists have observed that in many societies, violations of social norms are punished by ordinary (not official) individuals choosing to impose a costly penalty on the violator. Mahdi (1986), for example, shows the use of ostracism to punish norm violations among the Pathan Hill tribes in Afghanistan. In a cross-cultural survey, Boehm (1993) identifies several distributed mechanisms—ranging from social disapproval, criticism, and ridicule to disobedience and ultimately assassination—by which members of small-scale autonomous communities maintain egalitarian relationships and a lack of authoritative leadership by punishing those who attempt to dominate others. Wiessner (2005) documents the role of criticism, put-downs, pantomimes, mocking, complaints, and (infrequently) violence in norm enforcement among the Ju/'hoansi Bushmen of northeastern Namibia. Behavioral economists, in experiments conducted with students in university labs (Fehr & Gachter 2002; Fehr & Fischbacher 2004) and with individuals in a diverse set of populations in Africa, Asia, Oceania, South, and North America (Henrich

et al. 2006) have demonstrated a widespread willingness among humans to incur costs in order to punish those who violate norms.

Behavioral economists have suggested that altruistic punishment is explained by direct preferences over the behavior, payoffs or strategies of others (e.g., Levine 1998; Fehr & Fischbacher 2004). Fehr & Gachter (2002) suggest that altruistic punishment behavior is mediated by negative emotions such as anger toward rule violators. Evolutionary game theorists, however, have emphasized that it is challenging to explain how preferences for collective punishment—whether biological or cultural—could have evolved. Third-party punishment presents a free-rider problem. Punishment is costly to the punisher. The benefits that flow from punishment—inducing individuals to avoid violating social welfare-enhancing norms—are, however, collective goods enjoyed by punishers and non-punishers alike. Consequently, non-punishers enjoy higher fitness in a population with punishers and thus selection will favor non-punishers. Boyd & Richerson (1992) show that selection can favor third-party punishment strategies if such strategies also include punishment of non-punishers. This is an approach that is rooted in the concept of a subgame perfect equilibrium, and is the approach used, for example, by Milgrom, North, & Weingast (1990) to support an equilibrium in which cheating on contracts is deterred by an information sharing institution—which they call the Law Merchant—that coordinates collective punishment in a community of traders.[16] Bowles & Gintis (2004) use simulations to show that a stable population of strong reciprocators—individuals who incur personal costs to punish norm violations and generate group benefits—can emerge in a community that also includes those who violate norms and those who adhere to norms but fail to punish.

Boyd, Gintis, & Bowles (2010) present an evolutionary model that captures many of the same elements of collective punishment that we consider here. They presume, as we do, that cost-effective punishment requires multiple agents to decide to punish simultaneously; in particular, they assume increasing returns to punishment such that the cost of punishment falls as the number of punishers increases. At some threshold τ, given the (endogenous) likelihood of being in a group that has $\tau + 1$ punishers, punishment promotes the fitness of punishers. Importantly, their model allows punishers, as we do, to signal their

16 Greif (1994) proposes that cultural beliefs that include an expectation of collective punishment, together with cultural mechanisms that share information and coordinate expectations about what constitutes punishable behavior, can support a subgame perfect equilibrium in the absence of formal and centralized legal penalties. Subgame perfection in Greif's model of the Maghribi traders in the 11th Century (as in a version of Milgrom, North, & Weingast's [1990] model of the medieval Law Merchant) is achieved, however, because punishment is not costly for the punisher. A merchant in Greif's model is strictly better off punishing, by refusing to hire, an agent who has cheated a previous merchant in the past because the cheater will, in equilibrium, cheat the new merchant as well.

inclination to punish and thus save the costs of punishment if there are not enough other punishers around. They demonstrate that in such an environment, a population with punishers and non-punishers can be evolutionarily stable. Moreover, in a key overlap with our model, they demonstrate that in the stable state, the population of punishers will be such that there are likely to be just enough $(\tau + 1)$ punishers in a group, but no more, to make punishment worthwhile.

We add to this literature on collective punishment by providing another account of the incentive to participate in costly punishment. We suggest that even with standard materialistic preferences—no preferences directly over other's norm violations or heritable punishment strategies—there is an incentive to punish in order to communicate private information about the willingness to participate in supporting an equilibrium with coordinated punishment. Moreover, we demonstrate how these incentives can be harnessed by an institution that displays many of the characteristics we conventionally associate with law. Our model thus connects the literature on collective punishment and the evolution of cooperation to the analysis of the institutions that support distinctively legal (and distinctively human) order.

5. CONCLUSION

We began with the question, what is law? We do not have a complete answer to this question. Undoubtedly, law does involve in many cases the imposition of centralized coercive force and in order for the exercise of such force to constitute governance by law and not arbitrary power, law must have certain characteristics. Our answer to the question, what is law, is therefore that law can be a system of distinctive reasoning used to classify conduct as right or wrong that serves to incentivize and coordinate distributed agents in delivering punishments to deter wrongdoing. Our model is based exclusively on distributed enforcement that achieves effective deterrence of conduct that is deliberately classified as wrongful. It demonstrates that legal order can be sustained by a third-party institution that possesses many of the features that we intuitively associate with the concept or the rule of law: generality, universality, abstract and impersonal reasoning, open and public processes, stability, prospectivity, and clarity. The approach therefore makes a contribution to three literatures. Law and economics treats law like any other constraint and PPT treats law like any other means of chosing policy. Neither law and economics nor positive political theory of the law have an explanation for why law is characterized by distinctively legal attributes. Legal philosophy, on the other hand, has long debated what is distinctive about legal obligations but has not explored an

explanation for how the distinctive attributes of law arise or are sustained. Our model addresses all three of these lacunae.

We have shown that centralized coercion need not be present for legal order to emerge. (We provide examples of settings in which this is evident in our companion paper, Hadfield & Weingast [2011a].) As we show, decentralized enforcement is a possible substitute for centralized enforcement. Moreover, in many and perhaps all countries subject to law and rule of law, both third-party mechanisms are likely to be present.

Our focus on decentralized collective enforcement also brings to the fore a characteristic of legal order that is not as frequently emphasized, namely the role for an authoritative steward of a common logic that coordinates punishment by providing a system of unique classification. We demonstrate that these features of legal order can serve to solve the two key problems facing a community that seeks to deter wrongful conduct through decentralized collective punishment: they help to coordinate punishment decisions and to provide the incentive to incur the personal costs associated with punishment that benefits the larger group. Our positive analysis thus adds a new dimension to our understanding of the normative features of legal order. Most of the existing literature in legal theory looks to normative accounts of these normative characteristics. Open courts, impersonal reasoning and generality, for example, are frequently understood in terms of limits imposed by moral or political theory on the exercise of power by (particularly democratic) governments. We do not discount these normative limits, but we expand the understanding of these limits by showing how they may (also) be rooted in the positive or practical constraints on achieving stable equilibrium order based on law.[17]

From a positive perspective, our model sheds important light on fundamental questions of how, when and why distinctively legal order emerges in human societies. We have provided only one example of a legal order and the characteristics that serve to support that order. Our model is a very simple one. But our framework suggests several conjectures and avenues for further research. Here, we consider four key simplifications of our model.

First, our model does not consider how agents—buyers or sellers—will respond to ambiguity in classification. This is why our analysis can only be read to show that an unambiguous classification system is sufficient to support a form of legal order. We have provided intuition for why a less ambiguous classification system would be more likely to emerge in equilibrium, but we have not shown the extent to which ambiguity is disruptive of legal order equilibrium.

17 For an example of how the positive model sheds light on Rawls's normative theory of public reason, for example, see Hadfield & Macedo 2012.

A key extension of the model, then, is to explore the impact of variance and cost in classification. As we explore in Hadfield & Weingast (2011a), the emphasis we see in a wide variety of settings on the establishment of an authoritative steward with the capacity to render unique classifications of conduct suggests that the tolerance for ambiguity in a legal order—at least one based on decentralized enforcement—is low. This is not to say that a system cannot tolerate some ambiguity, and a more general model that allowed for noise in a classification system would help to determine how much is too much. This has implications for important policy questions in legal design, such as the balance between open-textured and plain meaning approaches to interpretation of legal documents, the relative values of certainty and flexibility in legal decisionmaking, and the extent of professional or hierachical control over the provision of legal advice.

Second, we have not modeled the supply or selection of the institution that coordinates equilibrium. But particularly in light of our emphasis on the decisionmaking attributes of the institution, such as its commitment to generality, impersonal reasoning and open process, it will be critical to explore the conditions under which an institution can be expected to conduct itself in this way. Here, our emphasis on decentralized enforcement and the need for R to offer benefits to the agents who participate in punishment, suggests new considerations—moving beyond the conventionally normative analysis of the duties of public officers such as judges to uphold the values of neutrality and openness. An institution that depends on the participation of citizens to achieve effectiveness faces incentives, as we have shown, to develop credible methods for ensuring desirable attributes such as impersonality and stable, public reasoning. Moreover, an environment that provides choice over alternative classification systems—such as existed in medieval Europe and in many other settings prior to the emergence of the nation state—creates the conditions for competition among institutions.[18]

Third, our model does not address a key challenge for collective punishment, namely the problem of free riding. We have only considered a setting in which participation by both potential victims in punishment is essential to effective deterrence. In a model with a larger number of agents, we would expect that even if multiple agents must participate for punishment to be effective, it will not generally be the case that *all* agents must punish all wrongs. This sets up the incentive for free riding on the punishment efforts of others, which conceivably

18 There is a significant literature on competition between states that supply regulatory regimes in corporate law, for example. See Hadfield & Talley (2006) for a discussion of this literature and its extension to competition between private providers. Hadfield (2012) discusses the emerging role for private production of legal systems in globalized settings.

could destroy a deterrence equilibrium. It is important to note, however, that the problem of free riding is not the same in our model as in most models of collective punishment. In our model, the incentive to punish is grounded in incentives to communicate information to others about the continued acceptability of the coordinating institution. A free rider in our model would not get a complete free ride: a failure to punish would come at the cost of causing other agents to downgrade their beliefs about the continued viability of a common logic. We conjecture that, particularly if only a subset of agents will be in a position to punish any particular violation, relaxing the constraint that all agents must punish will not destroy completely the potential for a deterrence equilibrium. Following Boyd, Gintis, & Bowles (2010), we expect that we could still demonstrate the viability of deterrence equilibria in environments in which agents find themselves facing the decision to punish or not in groups small enough that individual actions have a perceptible impact on beliefs about the likelihood of effective coordination in the future.

The risk of free riding as communities grow larger brings us to our fourth and perhaps most important modeling choice. We assumed that enforcement is exclusively achieved through a decentralized enforcement mechanism. Environments in which enforcement is decentralized are not hard to find, particularly prior to the emergence of the nation state (Hadfield & Weingast 2011a). But the problem of free-riding may well be a key reason for the state's consolidation of enforcement into a centrally controlled authority with a monopoly over legitimate coercive force. Here, however, our model suggests an intriguing hypothesis. We have shown a link between the normative characteristics frequently associated with the desirable attributes of "governance by law, not men" and the problem of coordinating and incentivizing collective participation in punishment. An institution that hopes to achieve effective legal order in this setting is constrained to ensure that its system is general, open, stable, impersonal, and so on. This suggests the possibility that a regime that relies on centralized coercive force is not similarly constrained. We wonder: does a shift to centralized enforcement come with a shift away from the rule of law? Or put differently: can any system that relies exclusively on centralized coercive enforcement be classified as a legal order? Or does it tend to (or necessarily) shift into a tyrranical or dictatorial order? We suspect that any order we would want to identify as legal must rely at least to some extent, and perhaps to a considerable extent, on decentralized enforcement. This might be true because a regime that depends exclusively on centralized punishment must expend exponentially increasing resources to manage a system of detection and punishment (exponential because delegation of these tasks to employees of the state requires enforcement of the rules governing these enforcers,) or rely on extraordinary, and disproportionate, penalties to compensate for only

probabilistic detection (Becker 1968). Our model suggests an additional reason to expect that reliance on exclusively centralized enforcement might be inconsistent with legal order: by relaxing the incentive constraint, a system of centralized legal enforcement is free to enforce rules that are indistinguishable from dictatorial fiat.

REFERENCES

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Basu, Kaushik. 2000. *Prelude to Political Economy: A Study of the Social and Political Foundations of Economics*. New York: Oxford University Press.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *J. Polit. Econ.* **76**, 169–217.

Binmore, Kenneth G. 1994. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

———. 1998. *Game Theory and the Social Contract: Just Playing*. Cambridge, MA: MIT Press.

Boehm, Christopher. 1993. Egalitarian Behavior and Reverse Dominance Hierarchy. *Current Anthropol.* **34**, 227–240.

Bowles, Samuel, & Herbert Gintis. 2004. The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations. *Theor. Population Biol.* **65**, 17–28.

Boyd, Robert, Harold Gintis, & Samuel Bowles. 2010. Coordinated Punishment of Defectors Sustains Cooperation and can Profilerate When Rare. *Science* **328**, 617–620.

Boyd, Robert, & Peter J. Richerson. 1992. Puishment Allows the Evolution of Cooperation (or anything else) in Sizable Groups. *Ethol. Sociobiol.* **13**, 171–195.

Bozovic, Iva, & Gillian K. Hadfield. 2012. Scaffolding: Using formal contracts to build informal relations in support of innovation, Manuscript, available at http://works.bepress.com/ghadfield.

Calvert, Randall, & James Johnson 1999. Interpretation and coordination in constitutional politics. In E. Hauser, & J. Wasilewski (*eds.*) *Lessons in Democracy*. Rochester: University of Rochester Press, pp. 99–138.

Cooter, Robert D. 1998. Expressive Law and Economics. *J. Legal Stud.* **27**, 585–608.

Crawford, Vincent P., & Hans Haller. 1990. Learning How to Cooperate: Optimal Play in Repeated Coordination Games. *Econometrica* **58**, 571–595.

Dixit, Avinash K. 2006. *Lawlessness and Economics: Alternative Modes of Governance*. New York: Oxford University Press.

Elkins, Zachary, Tom Ginsburg, & James Melton. 2009. *The Endurance of National Constitutions*. Cambridge, MA: Cambridge University Press.

Ellickson, Robert. 1994. *Order without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Fearon, James D. 2011. Self-Enforcing Democracy. *Quart J. Econ.* **126**, 1661–1708.

Fehr, Ernst, & Urs Fischbacher. 2004. Third-Party Punishment and Social Norms. *Evol. Hum. Behav.* **25**, 63–87.

Fehr, Ernst, & Simon Gachter. 2002. Altruistic Punishment in Humans. *Nature* **415**, 137–140.

Finnis, John M. 1980. *Natural Law and Natural Rights*. Oxford: Oxford University Press.

———. 1989. Law as Co-Ordination. *Ratio Juris* **2**, 97–104.

Fudenberg, Drew, & Eric Maskin. 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* **54**, 533–554.

Fuller, Lon. 1964. *The Morality of Law*. New Haven: Yale University Press.

Gans, Chaim. 1981. The Normativity of Law and Its Coordinative Function. *Israel L. Rev.* **16**, 333–355.

Green, Les. 1983. Law, Co-ordination and the Common Good. *Oxford J. Legal Stud.* **3**, 299–324.

Greif, Avner. 1994. Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies. *J. Polit. Econ.* **102**, 912–950.

Hadfield, Gillian K. 2012. Legal Infrastructure and the New Economy. *I/S: J. L. Policy Inform. Soc.* **8**, 1–59.

Hadfield, Gillian K., & Stephen Macedo. 2012. Rational Reasonableness: Toward a Positive Theory of Public Reason. *J. L. Ethics Hum. Rights*, (Forthcoming).

Hadfield, Gillian K., & Eric Talley. 2006. On Public Versus Private Provision of Corporate Law. *J. L. Econ. Organ.* **22**, 414–441.

Hadfield, Gillian K., & Barry R. Weingast. 2011a. Law Without Coercion: Examining the Role of Law in Coordinating Collective Punishment, Manuscript, available at http://works.bepress.com/ghadfield.

———. 2011b. Endogenous Institutions: Law as a Coordinating Device, Manuscript, available at http://works.bepress.com/ghadfield.

Hardin, Russell. 1989. Why a constitution? In Bernard Grofman, & Donald A. Wittman (*eds.*) *The Federalist Papers and the New Institutionalism* New York: Agathon Press, pp. 100–120.

———. 2006. Constitutionalism. In Barry R. Weingast, & Donald A. Wittman (*eds.*) *The Oxford Handbook of Political Economy*. Oxford: Oxford University Press, pp. 289–311.

Hart, H. L. A. 1961/1997. *The Concept of Law. 2nd edition*. New York: Oxford University Press.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, & John Ziker. 2006. Costly Punishment Across Human Societies. *Science* **312**, 1767–1770.

Hong, Lu, & Scott E. Page. 2001. Problem Solving by Heterogeneous Agents. *J. Econ. Theor.* **97**, 123–163.

Hume, David. 1739–40/1978. A Treatise of Human Nature. In L.A. Selby-Bigge, & P.H. Nidditch (*eds.*) 2nd edition. Oxford: Oxford University Press.

Kornhauser, Lewis A. 2004. Governance Structures, Legal Systems, and the Concept of Law. *Chicago-Kent L. Rev.* **79**, 355–381.

Kramarz, F. 1996. Dynamic Focal Points in N-person Coordination Games. *Theory and Decision* **40**, 277–313.

Levine, David K. 1998. Modeling Altruism and Spitefulness in Experiments. *Rev. Econ. Dynam.* **1**, 593–622.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Mahdi, Niloufer Qasim. 1986. Pukhtunwali: Ostracism and Honor among the Pathan Hill Tribes. *Ethol. Sociobiol.* **7**, 295–304.

Mailath, George J., Stephen Morris, & Andrew Postlewaite. 2001. Laws and Authority. Manuscript. University of Pennsylvania.

Mailath, G. J., S. Morris, & A. Postlewaite. 2007. Maintaining Authority. Manuscript. University of Pennsylvania.

Marmor, Andrei. 1998. Legal Conventionalism. *Legal Theor.* **4**, 509–531.

———. 2009. *Social Conventions: From Language to Law*. Princeton: Princeton University Press.

McAdams, Richard H. 2000. A Focal Point Theory of Expressive Law. *Virginia L. Rev.* **86**, 1649–1729.

———. 2005. The Expressive Power of Adjudication. *Univ. Illinois L. Rev.* **2005**, 1043–1121.

Milgrom, Paul R., Douglass C. North, & Barry R. Weingast. 1990. The Role of Institutions in the Revival of Trade: the Medieval Law Merchant, Private Judges, and the Champagne Fairs. *Econ. Polit.* **2**, 1–23.

Myerson, Roger B. 2004. Justice, Institutions, and Multiple Equilibria. *Chicago J. Intl L.* **5**, 91–107.

Ordeshook, Peter. 1992. Constitutional Stability. *Constitut. Polit. Econ.* **3**, 137–175.

Postema, Gerald J. 1982. Coordination and Convention at the Foundation of Law. *J. Legal Stud.* **11**, 165–203.

Raz, Joseph. 1977. The Rule of Law and Its Virtue. *L. Quart. Rev.* **93**, 195.
Schelling, T. C. 1981. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Selten, Richard. 1975. Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *Intl J. Game Theor.* **4**, 25–55.
Sugden, Robert. 1996. *The Economics of Rights, Cooperation and Welfare. 2nd edition*. New York: Palgrave Macmillan.
Sunstein, Cass. 1996. The Expressive Function of Law. *Univ. Pennsylvania L. Rev.* **144**, 2021–2053.
Waldron, Jeremy. 2008. The Concept and the Rule of Law. *Georgia L. Rev.* **43**, 1–61.
Weingast, Barry R. 1997. The Political Foundations of Democracy and the Rule of Law. *Am. Polit. Sci. Rev.* **91**, 245–263.
Wiessner, Polly. 2005. Norm Enforcement Among the Ju/'Hoansi Bushmen: A Case of Strong Reciprocity? *Hum. Nat.* **16**, 115–145.