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

Gillian K. Hadfield

University of Southern California and

Center for Advanced Study in the Behavioral Sciences at Stanford University

Barry R. Weingast

Stanford University and Hoover Institution on War, Revolution and Peace*

March 2011

Forthcoming, Harvard Journal of Legal Analysis

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 paper 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 begin with a simple set of axioms about what counts as legal order. We then 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 demonstrate 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.

I. 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?

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.[1] 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 paper we initiate the project of filling the gap in law and economics and PPT by developing a rational choice model of legal order. We begin with a few starting premises about what constitutes "order" and what makes it "legal." The concept of order is relatively straightforward. We say that order exists when behavior follows fairly predictable patterns: people pay their bills or drive on the right hand side of the road or seek the assistance of a mediator or religious leader in resolving a family dispute, for example. Of course, many sorts of order exist and many different mechanisms produce or contribute to producing order, including biology, technology, prices and social norms. So what makes order distinctively legal?

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 which 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.

Our approach begins with a parsimonious set of axioms that we take to be necessary (but not necessarily sufficient) attributes of any order that would be reasonably identified as being distinctively legal. We then build a rational choice framework based on those axioms. Using this framework we identify further characteristics of legal order based on analysis of the conditions under which an equilibrium can be stabilized in our model.

Specifically, we start with the following three axioms: First, if order is legal, behavior is patterned on a normative classification of behavior (such as "wrongful/not wrongful" or "punishable/not punishable"); in particular, behavior tends more frequently to be not wrongful or not punishable than wrongful or punishable. By normative we mean that the classification reflects an evaluative judgment by some agent(s); an action might be classified as wrongful or punishable, for example, because some agent(s) prefer that the action not be undertaken. The classification is not, in this sense, purely descriptive. A classification of pricing behavior in terms of the likelihood that the behavior will drive out competitors, for example, is purely descriptive; a classification of pricing behavior as illegal monopolization is normative. Legal order, we take it, is always normative. Note that our first axiom applies to any order based on social norms.

We take as our second axiom the assumption that if order is legal, the content of the normative classification is the product of deliberate choice by an identifiable entity. Our second axiom thus distinguishes legal order from spontaneous social order (Hayek 1960; Sugden 2005). Whereas the classifications we observe in spontaneous social order are empirical–the result in fact of social interactions, including the formation of beliefs–the classifications in a

legal order are attributable to a deliberate designation generated by an identifiable entity. Legal order is thus a vehicle for deliberate efforts to shape the adaptation of behavior to changes in the environment.

Our third axiom concerns the nature of the mechanism employed in a legal order that induces agents to choose not wrongful over wrongful actions. Specifically, we presume that when order is properly identified as legal, wrongful actions carry a penalty of some sort. Ordinary uses of the term law generally refer to penalty systems and we take as our domain the analysis of how systems based on penalties can generate legal order.[2]

Our third axiom allows for a range of motivations to avoid wrongful behavior, including: to avoid the discomfort of a social penalty, such as the disapproval of others; or to avoid a material consequence, such as when wrongful behavior is met with termination of a valuable relationship, a collective boycott, ostracism, or economic or physical retaliation.[3] Penalties may also be a product of internal reflection, as when individuals who prefer to act morally perceive an action as one that a conscientious person ought not to take; or psychological, such as when people feel guilty about violating the injunctions or expectations of others. Our use of the term wrongful suggests these internal attitudes towards law. But we do not presume that people have inherent preferences to avoid wrongful actions.

We emphasize that our third axiom 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 exclusively by decentralized mechanisms. This possibility marks a major departure from the implicit definition of law employed in most economic and positive political theory. As Dixit (2004, 3) observes, "conventional economic theory . . . assumes that the state has a monopoly over the use of coercion." Ellickson (1991, 127) defines law as rules that are enforced by governments rather than social forces.

By decentralized enforcement, we mean that the imposition of penalties is the result of individual decisionmaking among ordinary agents, not the decisionmaking of official actors

such as police or judges. 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 & Hamilton 1981), 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 2011), 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 starting axioms 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 are seeking 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 (that is, 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.

In this paper, we show that an equilibrium legal order can be secured in a setting in which penalties are delivered exclusively by decentralized collective punishment. 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 Section IV, below.) 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 and hence produces behavior that satisfies our first axiom for a legal order–requires coordination among multiple agents who must make simultaneous decisions 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 system of public and impersonal reasoning under what we call the *authoritative stewardship* of a third-party institution.

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. Instead, we focus on the incentive to participate in costly punishment as a basic problem to be solved 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 (Heinrich et al 2006) and 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 IV, below.)

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, in which the participation of all agents is necessary to 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 are attributes that secure, in our model, the incentive of individual agents to participate in collective punishment. Generality 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 collective punishment to 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 with 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 an account of the attributes of legal order that many intuitively associate with law–generality, impersonality, stability, openness, etc.–with 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 paper is organized as follows. Section II presents the model and Section III extracts the attributes of the equilibrium institution that generates legal order in that model. In Section IV, we relate our approach to the existing literature. Section V provides some concluding observations.

II. Model

Assume there are two infinitely-lived buyers, A & B who in each period $t = 1, ..., \infty$ purchase a good from an infinitely-lived seller.[4] 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 $X = (t, x_1, x_2, ...x_n)$ of factors with $x_1 \in \{A, B\}$ indicating the identity of the buyer. For each buyer j, let $\mathfrak{X}^j$ represent the set of n-tuples X in which the identity of the buyer $x_1 = j$.

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.

In each period and for each buyer, with probability θ, the seller has an opportunity to engage in a performance that is "wrongful" in some way. For simplicity we assume that any wrongful performance allows the seller to avoid the cost c. A key feature of this model is that we pay close attention to the inherent ambiguity of what it means for the seller's performance to be wrongful. A performance is deemed wrongful by a system of reasoning–a set of principles and procedures–or what we will call a *logic*. Formally a logic

maps the potentially very large set of all possible X vectors into a binary $\{0,1\}$classification of "wrongful" and "not wrongful."

Each buyer possesses an *idiosyncratic logic,* $I^j : \mathfrak{X}^j \to \{0,1\}$, to assess whether the seller's performance of a contract with that buyer is wrongful, that is, whether the buyer believes that the seller's contract obligated a delivery different in some way from the one performed by the seller. A performance might be judged by the buyer to be wrongful, for example, if fewer units or a lower quality of the good is delivered than the buyer expected, or if goods are delivered at a time or place different from the buyer's expectations, or if delivery is tendered subject to conditions the buyer did not believe to be included in the contract. (We will sometimes use the phrase "wrongful delivery" to mean any wrongful performance, including one in which no "delivery" is made at all.) This logic represents the buyer's assessment of the content of its own deals struck with the seller and the terms on which the buyer understands the seller to be obligated to deliver. Observe that the classification depends on t: the buyer's assessment of whether a performance is wrongful or not can change over time. For example, the buyer may have no preferences over packaging materials early on but may develop preferences (and hence contractual expectations) because of changes in warehouse practices, environmental regulations or consumer preferences. For ease of exposition, we make the simplifying assumption that any performance that the buyer judges to be wrongful according to I has no value to the buyer and yields the buyer a net payoff of $-P$.

Each buyer's idiosyncratic logic is not accessible to others: others cannot (at reasonable cost) reproduce the buyer's logic to predict how the buyer will analyze and hence categorize a given delivery failure. By designating this analysis as "idiosyncratic" we emphasize that the assessment 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). [5] Idiosyncrasy

may in part derive from private information and experience with factors that differs from others, but the buyer's logic is not 'information' in the sense of statements that can be unambiguously conveyed to others in a common language via a report or signal. The buyer's logic is a system of private reasoning to organize and analyze information and make judgments. Because people's situations and experiences differ, so too will their systems of judgment in ways that are not apparent to others. This is what makes idiosyncratic logic *ex ante* inaccessible to others.[6]

The seller's performances are observable to all. The seller keeps the payment P regardless of whether performance is judged wrongful or not. Buyers, after observing a wrongful delivery to any buyer, can choose to boycott the seller and obtain a payoff of 0 . Consider a seller's decision about whether to engage in a one-time wrongful delivery to a buyer when the seller anticipates that this might lead to a one-period boycott by one or both buyers. The 2-period profit for the seller who sells to and avoids wrongful delivery to both buyers is

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

The 2 period profit for the seller who sells to both buyers, makes a wrongful delivery to one buyer and expects to lose a next-period sale to that buyer only is

$$P+(P-c)+\delta(P-c).$$

The 2 period profit for the seller who makes a wrongful delivery to a buyer and expects to lose 2 future sales is

$$P+(P-c).$$

The seller will not be deterred by a lone one-period boycotter if

$$c>\frac{\delta}{(1+\delta)}P.$$

The seller will be deterred by two one-period boycotters if

$$c < \frac{2\delta}{(1+2\delta)}P.$$

We call a boycott that deters the seller from taking the opportunity to engage in a wrongful performance an *effective boycott* and assume that a boycott is only effective if both buyers participate in the boycott.

Condition 1. *Effective boycotting requires both buyers to simultaneously boycott for one period following a wrongful performance:*

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

We assume that both buyers will purchase the good in period t even if they anticipate that a wrongful delivery will never result in an effective boycott. Expected profits in this case are

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

Both buyers will therefore continue to purchase despite the absence of any threat of effective boycotting if

$$\theta < \frac{V-P}{V}. \tag{2}$$

(This assumption makes the exposition simpler and eliminates the need to analyze the incentive for the seller to coordinate boycotts–something to be considered in future work.)

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. There are, however, a variety of third-party institutions capable of classifying performances as wrongful or not and articulating reasons for their classifications. We refer to such a classification system as a third-party logic. Examples of institutions supplying a third-party

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. In this paper we do not seek to explain how various institutions arise or how a choice is made among the range of third-party logics on offer; we leave that for future work. Our more modest goal is to identify institutional characteristics that are sufficient (some necessary) to support an equilibrium in which buyers coordinate boycotting effectively to deter wrongful performances.

Assume that at $t = 0$ a seller has made a delivery to a buyer that would be classified as wrongful by at least one of the available third-party logics and that this has been observed by both buyers. In period 1, then, both buyers face a decision about whether to boycott the seller or not. Each buyer knows that the boycott will be effective only if the other buyer also boycotts. We assume that the buyers cannot get together to agree on a boycott; each decides independently whether to boycott.

As with any repeated coordination game, a subgame perfect Nash equilibrium to boycott can be supported by strategies under which buyers punish buyers who fail to participate in the boycott; punish buyers who fail to punish buyers who fail to participate in the boycott; and so on. (This is the type of strategy that supports the coordinated boycott equilibrium posited in Milgrom, North & Weingast (1990).) We exclude such strategies both on principle (it is hard to provide reasons external to the equilibrium for engaging in the strategy) and because of our assumption that each buyer possesses an idiosyncratic logic for classifying

non-deliveries as wrongful or not: there is no *a priori* unique common logic to determine when punishments should be delivered and when not. (Most game theoretic models of reputation and coordinated punishment, assume a unique common classification of actions as "cheating" or not.) We instead develop a model in which we can make it clear why an individual buyer might be willing to boycott based on the private benefit of a successful boycott, and we address explicitly how this incentive depends on the availability of a common logic with particular characteristics.

We allow a buyer when boycotting to make an announcement about the basis for the boycott. In particular, a boycotting buyer designates the logic under which it classifies the seller's performance as wrongful. Suppose, for example, that at $t = 1$ buyer A engages in a boycott and announces that it does so under a logic R. (The mnemonic is R for "reasoning" but it is important to remember that R is also a package of institutional characteristics such as the process for classification.) A's announcement can be interpreted as an endorsement of R although we do not ascribe to A any inherent preferences (such as cultural beliefs) directly over R: A does not derive utility directly from the adoption or endorsement of R by itself or others, but only from the consequences of choices made from following (or not following) R. For ease of exposition we will refer to a performance classified as wrongful by R as "R-wrongful."

Clearly the only reason for A to engage in this behavior is if by doing so A changes the likelihood that the seller makes a wrongful delivery to A in the future. A solitary boycott, however, does not change the seller's expected payoff sufficiently to deter a wrongful delivery. Thus A's decision (including the announcement of R) must be premised on the impact of its actions on the seller's expectation of a two-buyer, effective, boycott in the future. This, in turn, depends on the impact of A's choice on B's beliefs and actions.

We start by examining B's decision in period 2, following the observation at $t = 1$ that A has boycotted and designated R. To see how the intuition of our model works, suppose that A announced an R with the attribute that $R(X) = I^B(X)$ for all $X \in \mathfrak{X}^B$. That is,

suppose A designated the logic underlying its boycott as the equivalent of B's idiosyncratic logic for cases involving B. And suppose that B believes that A will continue in the future to boycott any R-wrongful performance if A observes B to join in the boycott. (We consider the consistency of this belief in the proof of Proposition 1, below.) What is B's incentive to join in the boycott in period 2? Provided the seller also believes that buyers who participate in an effective boycott will continue to boycott R-wrongful performances in the future, by joining A's boycott, B's choice to boycott under R changes the seller's beliefs about the likelihood of an effective boycott in the future. This expectation induces the seller to avoid performances that can trigger a boycott, that is R-wrongful performances. Because, in this example, R coincides with B's idiosyncratic assessment of when the seller has violated the contract with B, this implies that B secures the elimination of performances it judges to be wrongful, starting in period 3. Formally, joining the boycott in period 2 will then be worth it to B if

$$\sum_{t=3}^{t=\infty} \delta^{t-2} (V - P) > \sum_{t=2}^{t=\infty} \delta^{t-2} ((1-\theta)V - P). \tag{3}$$

Rewriting 3 we have the condition that, assuming that both A and the seller believe that once A and B have coordinated on R they will continue to boycott R-wrongful performances in the future, B will join the boycott if

$$\theta > (1-\delta)\frac{V-P}{V}. \tag{4}$$

That is, B will join the boycott provided that θ, the risk of performances classified as wrongful under B's idiosyncratic logic, is sufficiently great or the discount factor δ is sufficiently close to 1. This result does not require that B be the victim of the original wrongful performance. B's incentive to boycott in response to a wrong to A arises because under the logic designated by A and with the assumption that coordination to boycott wrongs under R will persist once established, all R-wrongful performances are deterred. For B, this includes all potential

wrongs done to B. This analysis demonstrates the potential for A to designate a logic and demonstrate a willingness to boycott so as to induce the expectation of a coordinated boycott in response to wrongs in the future.

We have no reason, of course, to believe that A will in general want to coordinate boycotts for I^B-wrongful performances unless the logic that does so also deters sufficiently often performances that A classifies as wrongful under I^A. Moreover, by assumption, B's logic is presumed to be idiosyncratic and inaccessible to others. Thus even if A wanted to, it is not possible for A to deliberately designate a logic that replicates I^B on $\mathfrak{X}^B$. Nor is it possible for the seller to condition future delivery decisions on whether or not they will trigger a boycott coordinated by judgments reached using I^B.

We do not take on in this paper the task of modeling the selection of R. Instead we demonstrate characteristics of R that will support an equilibrium in which the buyers are coordinated in effective boycotting and the seller is able to predict the effect of alternative performances on the likelihood of a boycott. We proceed by characterizing an equilibrium and then examining the features of R as an institution that support the existence of that equilibrium.

We will develop the implications of the model for the characteristics of R as an institution more fully in Section IV, below. For now we proceed to develop the formal model by assuming that A can designate an $R : \mathfrak{X}^A \cup \mathfrak{X}^B \rightarrow \{0, 1\}$ that possesses the minimal qualities of being publicly (and freely) accessible and stable: all players can access the logic at zero cost to determine how it would classify a performance vector X and all players expect R's classifications to remain unchanged in all future periods. We will call this a stable *common logic*.

Let r be a measure of the expected *convergence* between a buyer's idiosyncratic logic and the common logic R. Formally, let r_t^j be buyer j's estimate in period t of the likelihood that given an opportunity for wrongful performance in the future, R and I^j will both classify performance as wrongful. We abstract here from the implications of an R that makes type

1 errors (finding a performance wrongful that is not judged by the buyer to be wrongful) by assuming that if a buyer does not judge a performance to be wrongful the other buyer does not learn of it and so it cannot trigger an effective boycott. Thus we do not address the potential for strategic behavior on the part of the buyer who wishes to extract value from a seller by threatening to induce a boycott under R even when a performance conforms to the buyer's original expectations.

Recall that I^j can change over time, implying that buyer j may adjust this estimate over time. To keep things simple, we assume that, conditional on information available in period t, a buyer uses the same estimate, r_t^j, for all periods $t+1$. To reduce notation, we will suppress the subscript t for estimates made in the first two periods and assume that the estimate does not change over this time period.

Evaluated as of period 1, the expected payoff to buyer j if R, as of period 3, coordinates the expectations of both buyers and the seller that an R-wrongful performance will be met with an effective boycott is then given by

$$\sum_{t=3}^{t=\infty} \delta^{t-1}(1-(1-r^j)\theta)V - P). \tag{5}$$

The expression in brackets preceding V in (5) reflects buyer j's belief about that probability that, with the threat of an effective boycott in response to R-wrongful performances, it will enjoy performance with value V. It does so in all cases except where the seller has an opportunity to engage in a performance that j's idiosyncratic logic classifies as wrongful but R does not; in those cases the seller delivers a performance that the buyer classifies as wrongful, with a value to the buyer of 0, producing a net payoff for the buyer of $-P$. Note that if $r^j = 1$, meaning that R is fully convergent with I^j, then the buyer always receives performance with value V, producing a net payoff of $V - P$ in each period.

It is clear from (5) that A would prefer to designate a common logic R that is as convergent as possible with I^A, that is, with r^A close to 1. But A's effort to coordinate the

market on R will only be successful if B chooses to join the boycott under R. B will only join the boycott if B expects to do better in a coordinated equilibrium that deters R-wrongful performances than in the uncoordinated equilibrium. This requires r^B sufficiently high to satisfy B. Let p be A's belief about the probability that B will join the boycott in period 2. If A is willing to boycott in period 1, then, it must be that

$$(6) \qquad \begin{aligned} & p\sum_{t=3}^{t=\infty}\delta^{t-1}\left[(1-(1-r^A)\theta)V-P\right]+(1-p)\sum_{t=3}^{t=\infty}\delta^{t-1}\left[(1-\theta)V-P\right] \\ > \; & \sum_{t=1}^{t=\infty}\delta^{t-1}((1-\theta)V-P). \end{aligned}$$

We will say that R is *sufficiently convergent* with I^j if r^j is high enough, given any risk that the other buyer will not join the boycott, to make j better off under an equilibrium coordinated on R than not. Rewriting (6) above, this gives us the following necessary condition for A's participation in the early stages of a sequence leading to an equilibrium coordinated by R:

Condition 2. *A necessary condition for buyer A to be willing to begin a boycott in period 1 and designate a logic R as the basis for the boycott is that, conditional on A's beliefs in period 1, R is sufficiently convergent with I^A:*

$$(7) \qquad \begin{aligned} r^A \; & \geq \; \frac{1}{p}\frac{(1-\delta^2)}{\delta^2}\left[\frac{(1-\theta)}{\theta}-\frac{P}{\theta V}\right] \\ & \equiv \; r^1 \end{aligned}$$

Satisfying Condition 2 requires that $0 \leq r^1 \leq 1$. The right hand side of (7) is always positive given the assumption that buyers are willing to purchase even without deterrence. Then for given p, $r^1 \leq 1$ can be ensured by choosing δ close enough to 1. Note that the greater is A's confidence that R will be sufficiently convergent for B, the more likely it is that A will be willing to risk boycotting unilaterally in an effort to coordinate with B.

For B to be willing to join the boycott in period 2, it must be that

$$\sum_{t=3}^{t=\infty} \delta^{t-2}((1-(1-\theta+r^B)\theta)V-P) > \sum_{t=2}^{t=\infty} \delta^{t-2}((1-\theta)V-P). \tag{8}$$

Under the assumption that A will boycott for two periods, B does not risk being a lone boycotter in period 2. Rewriting, this gives us the following necessary condition for B's willingness to join a boycott, on the assumption that this will result in an equilibrium coordinated by R:

Condition 3. *A necessary condition for buyer B to be willing to join a boycott in period 2 given the designation by A of R as the basis for the boycott is that R is sufficiently convergent with I^B:*

$$\begin{aligned} r^B &\geq \frac{(1-\delta)}{\delta}\left[\frac{(1-\theta)}{\theta} - \frac{P}{\theta V}\right] \\ &\equiv r^2 \end{aligned} \tag{9}$$

Satisfying Condition 3 requires that $0 \leq r^2 \leq 1$. Again, the right hand side of (9) is always positive given the assumption that buyers are willing to purchase even without deterrence. $r^2 \leq 1$ can then be assured for δ close enough to 1. Moreover, observe that

$$r^1 > r^2.$$

This implies that δ sufficiently close to 1 to ensure $r^1 < 1$ will also ensure $r^2 < 1$.

The requirement that $r^1 > r^2$ is intuitive: as the first mover who bears the cost of a longer boycott in order to test the unknown acceptability of R to B, A must enjoy a higher minimum return than B from the implementation of R. Second, B must bear the cost of a one-period boycott to signal the acceptability of R. Although this is a lower cost than A bears, it still imposes a constraint on the nature of R in equilibrium: A cannot establish an

equilibrium with R if R shows too little convergence with B's idiosyncratic logic; by moving first A cannot pull the equilibrium too close to its own idiosyncratic logic if that pulls it too far away from B's. Only if R is sufficiently convergent is a buyer better off with deterrence of R-wrongful performances than without.

Conditions 2 and 3 gives us necessary conditions for the emergence of an equilibrium coordinated by R, but not sufficient conditions. In order to establish an equilibrium we have to look at the beliefs that underlie the incentives of A and B to start or join a boycott organized on the basis of R and the beliefs of the seller. We turn to beliefs now.

Conditions 2 and 3 capture the decision problems for A and B, respectively, only if all players–both buyers and the seller–believe that if an effective boycott based on R is coordinated in period 2, then both buyers can be expected to boycott in any future R-wrongful performance. This is what produces the payoff starting in period 3 and continuing into the infinite future in which the buyer enjoys a lower rate of wrongful performances: sellers anticipate that an R-wrongful performance will lead to a 2-buyer boycott and will therefore (by Condition 1) be deterred from R-wrongful performances. Additionally, both buyers must believe that if either A fails to boycott in period 1 or B fails to boycott in period 2, or if either buyer fails to boycott an R-wrongful performance in the future, then no buyer will ever boycott R-wrongful performances in the future. This is what produces the payoff resulting from a decision not to initiate (A) or join (B) a boycott.

These beliefs can be sustained by the following reasoning. Recall that r_t^j is buyer j's estimate in period t of the likelihood that given an opportunity for wrongful performance in the future, R and I^j will classify performance in the same way. Because I^j can shift over time, leading to an update in the buyer's estimate r_t^j over the remaining (infinite) future, in any period t there is uncertainty about whether R continues to be sufficiently convergent for a particular buyer. Let ε_t be the belief held by all players in period t about the probability, conditional on an effective boycott having been organized by R, that R ceases to be sufficiently convergent for a particular buyer. Suppose that an effective boycott

has been observed in period 2 and that in period $\tau > 2$ the seller engages in an R-wrongful performance. Clearly if R is no longer sufficiently convergent for a buyer, that buyer will not boycott in period $\tau + 1$. (The precise meaning of sufficient convergence in this context is given in the proof of Proposition 1.) Consider buyer A and assume R is still sufficiently convergent with I^A_τ. Buyer A knows that buyer B entertains the prior belief ε_τ that R is no longer sufficiently convergent for buyer A. Consistent with the equilibrium strategy, if buyer A fails to boycott, buyer B will update this probability to infer that R is no longer sufficiently convergent for A. With that updated belief, buyer B will engage in no future boycotts. If buyer A does boycott, then buyer B will update its prior to infer that R is still sufficiently convergent. Buyer A is therefore better off boycotting than not in period $\tau + 1$ in order to prevent B from inferring that R is no longer capable of coordinating an effective boycott. The same reasoning applies to buyer B's decision to boycott in period $\tau + 1$.

We can now state our main result.

Proposition 1. *Given that Conditions 1, 2 and 3 are satisfied, the logic R and the following strategies and beliefs support a perfect Bayesian equilibrium in the repeated game such that beginning in period 3 all players will expect a coordinated boycott in response to R-wrongful performances and the seller will be deterred from R-wrongful performances so long as R remains, and is expected to remain, sufficiently convergent for both buyers: (1) Following an R-wrongful performance in period* 0*, A boycotts in periods 1 and 2 and announces R; (2) B boycotts in period 2 if A boycotted in period 1; (3) thereafter, if and only if an effective boycott was achieved in period 2, each buyer j engages in a next-period boycott whenever an R-wrongful performance occurs and R remains sufficiently convergent with I^j; (4) the seller exploits the opportunity to engage in a wrongful performance in every period unless and until an effective boycott in achieved in period 2; (5) all agents entertain the belief that for either buyer, I^j is sufficiently convergent with R if and only if the buyer boycotts as prescribed by the equilibrium strategies and believe that, conditional on an effective boycott in period 2, with probability ε_t, R remains sufficiently convergent to induce each buyer to participate in*

a boycott in period t.

Proof. See appendix.

III. Discussion

We have shown that a 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. It is an order because in the equilibrium, behavior has been effectively elicited to follow a designated pattern: buyers enter into contracts with sellers, pay them up front and receive deliveries of goods that a known common logic classifies as not wrongful. It is a legal order because it satisfies our three axioms: 1) order is based on a normative classification, with behavior classified as not wrongful favored over behavior classified as wrongful; 2) the order is conditioned on the particular content of the classification system deliberately chosen by a third-party institution, R; and (3) the avoidance of wrongful behavior is based on a system of punishment.

The order structured by R in our model possesses several additional attributes that are commonly associated with the existence of law or the rule of law. We consider these attributes in turn.

Generality.— Our model makes it plain that the logic R necessarily must be general in the particular sense that it addresses itself to the interests and situations of both buyers. Mathematically, this means that R must be defined over the set of circumstances that both A and B consider relevant: $R : \mathfrak{X}^A \cup \mathfrak{X}^B \to \{0, 1\}$. Generality derives from the

equilibrium requirement that A designate a common logic that is sufficiently convergent with B's idiosyncratic logic to attract B to participate in a coordinated boycott triggered by the application of the common logic. If, instead, A selects a logic that is too personalized, too focused on protecting A's interests alone, B will refuse to participate in the boycott.

In a deeper sense, the model can be interpreted implicitly to presume that R will be general in the sense of articulating its classification scheme in abstract rather than concrete terms. The reason is as follows. Our model presumes that the common logic designated by A is provided by a third-party institution and that in equilibrium R will be sufficiently convergent with the idiosyncratic logic of both buyers. We defined an idiosyncratic logic as one that is largely inaccessible, at reasonable cost, to other actors. This includes the actors who supply R. Although we have not modeled A's selection of R, or the process by which a set of available institutions is generated, given that both production and designation of R must be made without knowledge of the specifics of at least one buyer (B), the chances that R will emerge as an equilibrium will be maximized if R is expressed in abstract (general) terms that are more likely to accommodate the (unknown) specifics of B's idiosyncratic logic. If R is simply a classification of fact-specific circumstances (generated from past experiences among the actors who produce R, for example) rather than general descriptions of circumstances–if the classification of specific circumstances is not *generalizable*–then the likelihood that R will be assessed by B to be sufficiently convergent (the probability p in the model) will be relatively low in an environment with sufficient heterogeneity. Our assumption of idiosyncrasy is intended to capture this heterogeneity.

Similarly, although we have not modeled this explicitly, if B is unknown to A and the third-party institution, 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. This is another aspect of generality. 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 in the same way), we would conjecture

that, with little ability to predict the particular content of B's idiosyncratic logic, A can increase p 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 generality 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 generality in our model serves to support the punishment incentives of agents whose participation in punishment is necessary for effective deterrence. If we were to add a third buyer C, and continue to suppose that effective deterrence still only required a two-buyer boycott, 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 "general" 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 general with respect to C as well.

Clarity and Uniqueness.—The model assumes that it is common knowledge that once R has been established in equilibrium, all agents will agree on what constitutes an R-wrongful performance and hence all will 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 classfications–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 *authoritative stewardship* 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 in an established legal system in a complex environment there is usually a single Supreme Court, for example.

With the model we have presented, we can only claim to have shown that clarity and uniqueness are sufficient to support an equilibrium under R with effective deterrence. We would conjecture that equilibrium could be supported with some degree of 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 Proposition 1 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 wrongful deliveries 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.

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 (ε_t) arising from drift in idiosyncratic logic that opens up too wide a gap between R and I^j. 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. With common knowledge unique classification, there is only one consistent inference that can be drawn from buyer j's failure to boycott: R is not, or is no longer, expected to be 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. As with generality, although we have not modeled A's choice among an array of available institutions offering a common logic, it seems clear that the probability that A will succeed in establishing a deterrence equilibrium if A designates an institution that more effectively reduces ambiguity. Similarly, if institutions are competing for selection, the agents controlling a common logic will be more likely to secure selection of their institution if they more effectively achieve unique and clear classifications.

Impersonal Reasoning .—In the model as we have presented it–with only two buyers and a single seller–it may seem reasonable to suppose that equilibrium could be supported by the idiosyncratic logic of the institutional agent(s) supplying the classification services of R. The model would only require that the buyers or seller be able to query this institutional agent to learn the classification that would be made of any 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. 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 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.

Impersonal reasoning implies that the institution providing the logic 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–even presuming that A could communicate the content of its idiosyncratic logic and, further, even presuming that I^A and I^B are sufficiently convergent–A cannot just propose I^A –a logic it controls–as the basis for its boycott.

Neutrality and independence are routinely identified by analytical 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." In our model, 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. Neutrality reduces ambiguity.

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 logic to consult it in 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.[7] Our model suggests that a robust common logic is likely to be a form of *public reasoning* elaborated in an *open process* to which an interested party might introduce their private information and reasoning.

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, in a homogeneous world with shared and unambiguous classifications of all performances as "cheating" or not, there is no need at all for an external institution to provide a common logic; in such a world, we will find it easier to predict, as do Hume (1740) and Sugden (2005), the spontaneous emergence of norms to coordinate behavior. 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 appears to follow from our model because the assumption of idiosyncrasy suggests that the classifications reached by the logic must be *immanent*, not fully articulated in any form that can be consulted *ex ante* by all agents. (The idea of immanence will be recognizable to those schooled in the traditional legal concept that the common law is "found" not "made": it contains all of its principles even if they are not articulated until a specific case is adjudicated.[8]) 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. (In a world where delivery in 10 days 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.) Having no access to the idiosyncratic reasoning of individual buyers when it offers its logic as a candidate coordination device in period 1, 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, is unlikely to be (just) a dataset collecting classifications already reached by the logic; it is likely to contain the placeholders for dealing with as-yet-unimagined circumstances. Nor is it likely to be a complete prescription of how all possible circumstances would be classified by the logic. To do 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 immanent, 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 produce 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-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, although we are not modeling the selection of A as a strategic choice in our simple model, we would predict that A would be more likely to propose an R that is open to hearing from B and sufficiently public. Open process and public reasoning are likely to give B greater confidence that R will, in practice, converge sufficiently with I^B.

Stability.—All legal theorists emphasize that law must consist of relatively stable rules. Our model also assumes this. But conventional accounts of law, focused on the need to provide individuals with sufficient guidance that they can conform their conduct to law and so avoid punishment, imply a different timeframe for stability. In conventional accounts, a rule needs to be stable between the time an agent chooses an action and the time at which there is the potential for having that action judged and penalized under the rule. This is the timeframe the seller in our model cares about: rules need to be stable during a period, but from the seller's perspective they could change period to period.

Stability in our model, however, is also required to meet the requirements of the buyers.

They require stability over a much longer horizon than the seller does. The buyers must be able to anticipate in the early stages of the game (periods 1 and 2) that the logic they are evaluating as a potential coordinating device will retain its classifications over an infinite horizon. The model allows for some drift in the relationship between R and I^j over time, but equilibrium requires that the probability ε_t that drift leads to a sufficient divergence between R and I^j to destroy buyer j's incentive to continue to boycott R-wrongful performances be sufficiently small. That is, equilibrium requires sufficient stability to support incentives to boycott, not merely to provide a stable guide for seller behavior.

Prospectivity.—Compare the departure between our theory and conventional legal theory with respect to stability to the implications for prospectivity. Conventional legal theory, again on the basis of what a person requires in order to conform and avoid punishment, asserts that law must be prospective: the seller cannot condition behavior at the start of period N on a rule that is not available until the end of period N. But rules could change from period to period. In our model, buyers do not care about prospectivity except to the extent that they can predict that if they coordinate on R, the seller will be able to effectively condition its behavior on R. Thus our prospectivity requirement is addressed to the same need as that proposed in conventional theory.

IV. Relationship to the Literature

A. Philosophy of Law

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 paper. 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 understand our work to relate 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 eight characteristics: 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 and the rule of law in this

paper 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? What are the mechanisms by which law achieves order and how can these mechanisms be structured or modified to achieve particular ends? The definition of law that we work towards 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 ([1960] 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 spontaenous social order. Our contribution is to take this insight more firmly in the direction of social scientific, particularly rational choice, analysis.

B. Coordination Accounts of Law

There is a large literature in both social science and legal philosophy, going back to Hume (1739-40), exploring the idea that law plays a role in coordinating behavior.

In legal philosophy, coordination accounts have been largely spurred by Hart's (1960) 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], 2002), 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 (1975), Finnis (1980, 1989), Gans (1981), Marmor (1998, 2009) and Green (1983). Although this literature employs 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 and Melton 2008), so why do those few survive? Hardin (1989, 2006), following Hume (1739-40), argues that the central feature of constitutions is to provide coordination for citizens around various rules (see also Ordeshook 1992, Calvert & Johnson 1998). 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 (2006) 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 (2005) 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 (2004) 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 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 a 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 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 than 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 such as a legal institution or practice 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.[9] In our model, in contrast, the problem of coordination is one faced by agents who are potentially engaged in punishing primary behavior: responding to those who drive on the wrong side, take what is not theirs or adopt unconventional statutory readings. That is, we focus on the role of legal rules governing primary behavior in solving the coordination problem facing those who may participate in enforcing 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 only 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.

C. 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 maintained 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 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 towards 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 pun-

ishment 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 violations–are, however, 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 sub-game perfect equilibrium, and is the approach used, for example, by North, Milgrom & 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.[10] 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 and 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 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 stable state, the population of punishers will 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 a 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.

V. Conclusion

We began with the question, what is law? Our answer is that law is, at least in part, a system of distinctive reasoning used to classify conduct as right or wrong that serves to coordinate distributed agents in delivering punishments to deter wrongdoing. Our model demonstrates that a legal order based exclusively on distributed enforcement (of the type we model) that achieves effective deterrence of conduct that is deliberately classified as wrongful can be based on a third-party institution that possesses many of the features that we intuitively associate with the concept or the rule of law: generality, abstract and impersonal reasoning, open and public processes, stability, prospectivity and clarity. Our focus on 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 the logic the coordinates punishment by providing a system of unique classification. We demonstrate that that these features of legal order can serve to solve the two key problems facing a community that seeks to deter wrongful conduct through 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 our 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. (For an example of how the positive model sheds light on Rawls's normative theory of public reason, for example, see Hadfield & Macedo 2011.)

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 paper 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 equilibrium legal order. A key extension of the model, then, is to explore the impact of variance, and cost, in classification. As we explore in Hadfield & Weingast (2011), 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 the interpretation of legal documents, the relative values of cer-

tainty 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, R. 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.[11]

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 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 would conjecture that, particularly if we assume only for a smaller 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. As in Boyd, Gintis & Bowles (2010), we would 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. 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 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 (including voluntary compliance in our definition of decentralized enforcement.) This might be true because an exclusively centralized system of punishment must expend perhaps expoentially 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 that reliance on exclusively centralized enforcement might be inconsistent with legal order also because, by relaxing the incentive constraint, a system of centralized legal enforcement is free to enforce rules that are indistinguishable from dictatorial fiat.

VI. Appendix

Proof of Proposition 1. It is straightforward to see that if Conditions 1, 2, and 3 are satisfied, both buyers are better off boycotting as prescribed than not, and the seller is better off not engaging in R-wrongful performances, provided all players believe that any R-wrongful performance in any period $t > 2$ will be met with an effective boycott. Here we check that these beliefs are consistent with sequentially rational strategies for all players. Consider first the sequential rationality of buyer A's equilibrium strategy, which calls for A to boycott in the event of an R-wrongful performance in any period $t > 2$, provided R is still sufficiently convergent with I^A. Suppose in particular that the seller makes a one-time out-of-equilibrium move and engages in an R-wrongful performance to some buyer in period $\tau - 1$ where $\tau > 3$. In that event, given the proposed equilibrium beliefs, B would infer from A's failure to boycott in period τ that R is no longer sufficiently convergent with I^A. This implies that, expecting A not to boycott, B will also not boycott in any period $t > \tau$. Then A's expected payoff if deviating from the equilibrium strategy and not boycotting in peiod τ is

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

A's expected payoff if choosing to boycott in period $\tau+1$, given equilibrium beliefs and equilibrium strategies by B and the seller, is

$$(1-\varepsilon_\tau)\sum_{t=\tau+1}^{t=\infty}\delta^{t-\tau}\left[(1-\theta+r_\tau^A\theta)V-P\right]+\varepsilon_\tau\sum_{t=\tau+1}^{t=\infty}\delta^{t-\tau}\left[(1-\theta)V-P\right].$$

Then deviation from the equilibrium strategy is not optimal for A in period τ so long as

$$\begin{aligned} r_\tau^A &\geq \frac{1-\delta}{\delta}\frac{1}{1-\varepsilon_\tau}\left[\frac{(1-\theta)}{\theta}-\frac{P}{\theta V}\right] \\ &\equiv r^\tau. \end{aligned} \tag{10}$$

We then have that $r^\tau \leq r^2 < r^1$ for ε_τ sufficiently small. Thus, boycotting is optimal for A in period τ whenever an effective boycott based on R has been observed in period 2 provided R remains sufficiently convergent for A and R is expected to remain sufficiently convergent for B with high probability (low ε_t). The proof for B is identical as A and B are symmetric beginning in period 3.

We now check the seller's strategy and consider whether a one-time deviation from the equilibrium strategy in some period $\tau > 2$, making a wrongful delivery to one buyer, generates a higher payoff. A one-time one-buyer deviation generates a payoff from the wrongful performance of P and $P-c$ from the other buyer in period τ. A one-time deviation implies that the seller expects to sell and not perform wrongfully in all other periods and so we can consider only the two-period payoff to determine the seller's optimal decision in period τ. In period $\tau > 2$ after observing an effective boycott in period 2, the seller entertains the belief that with probability ε_τ, R will have ceased to be sufficiently convergent for each buyer j. The seller also expects, under equilibrium beliefs, that if it makes a wrongful delivery in period τ each buyer will boycott in period $\tau+1$ if and only if R remains sufficiently convergent for that buyer. This implies that the seller expects to sell to each buyer in period $\tau+1$ with probability ε_τ. These are independent events so the seller expects a payoff in period $\tau+1$ of $2\varepsilon(P-c)$ which is discounted by δ. Thus the seller who engages in a

one-time one-buyer wrongful performance in period τ expects a 2-period payoff of

$$2P - c + 2\delta\varepsilon(P - c).$$

(We assume that in all other periods, the seller continues to play the equilibrium strategy and avoid R-wrongful deliveries.) The expected payoff for proper performance is

$$2(1 + \delta)(P - c).$$

Thus the seller will be deterred from wrongful performance if

$$c < \frac{2\delta(1 - \varepsilon_\tau)}{1 + 2\delta(1 - \varepsilon_\tau)} P.$$

Given satisfaction of the original effective boycotting condition, Condition 1, the above condition will be satisfied for ε_τ sufficiently close to 0.

It remains to check that the proposed equilibrium beliefs are consistent with the equilibrium strategies. The belief system calls for a buyer or the seller to infer that R is no longer sufficiently convergent for j if and only if j fails to boycott as prescribed. As shown above, because boycotting is the optimal strategy for j if and only if R is sufficiently convergent with I^j, this inference is consistent. The belief system also calls for all players to believe that in any period $t > 2$, following the observation of an effective boycott in period 2, a buyer will boycott in response to an R-wrongful performance with probability $(1 - \varepsilon_t)$. ε_t is the common belief held by all players about the likelihood that, conditional on R supporting an effective boycott in period 2, buyer j has downgraded its belief r_t^j sufficiently to destroy j's benefit from deterrence of R-wrongful performances. Given that the optimal strategy is to boycott so long as R remains sufficiently convergent and not otherwise, ε_t is therefore also the likelihood that a buyer in period t will fail to boycott an R-wrongful performance.

REFERENCES

Axelrod, R. and W. D. Hamilton (1981). The evolution of cooperation. *Science 211*, 1390–1396.

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

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*, 169–217.

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

Binmore, K. G. (1998). *Game Theory and the Social Contract: Just playing*. Cambridge, MA: MIT Press.

Boehm, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Current Anthropology 34*, 227–240.

Bowles, S. and H. Gintis (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology 65*, 17–28.

Boyd, R., H. Gintis, and S. Bowles (2010). Coordinated punishment of defectors sustains cooperation and can profilerate when rare. *Science 328*, 617–620.

Boyd, R. and P. J. Richerson (1992). Puishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology 13*, 171–195.

Calvert, R. L. and J. Johnson (1999). Interpretation and coordination in constitutional politics. In E. Hauser and J. Wasilewski (Eds.), *Lessons in Democracy*. University of Rochester Press.

Cooter, R. (1998). Expressive law and economics. *Journal of Legal Studies 27*, 585–608.

Crawford, V. P. and H. Haller (1990). Learning how to cooperate: Optimal play in repeated coordination games. *Econometrica 58*, 571–595.

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

Elkins, Z., T. Ginsburg, and J. Melton (2009). *The Endurance of National Constitutions*. Cambridge: Cambridge University Press.

Fehr, E. and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*, 63–87.

Fehr, E. and S. Gachter (2002). Altruistic punishment in humans. *Nature 415*, 137–140.

Finnis, J. M. (1980). *Natural Law and Natural Rights*. Oxford: Oxford University Press.

Finnis, J. M. (1989). Law as co-ordination. *Ratio Juris 2*, 97–104.

Fuller, L. (1964). *The Morality of Law*. New Haven: Yale University Press.

Gans, C. (1981). The normativity of law and its coordinative function. *Israel Law Review 16*, 333–355.

Green, L. (1983). Law, co-ordination and the common good. *Oxford Journal of Legal Studies 3*, 299–324.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy 102*, 912–950.

Hadfield, G. K. (2010). Law for a flat world: Legal infrastructure and the new economy. Available at ssrn.com.

Hadfield, G. K. and S. Macedo (2011). Rational reasonableness: Toward a positive theory of public reason. Available at ssrn.com.

Hadfield, G. K. and E. Talley (2006). On public versus private provision of corporate law. *Journal of Law, Economics and Organization 22*, 414–441.

Hadfield, G. K. and B. R. Weingast (2011). Coordinating collective punishment: A coordination account of legal order. Available at ssrn.com.

Hart, H. L. A. (1997). *The Concept of Law* (2 ed.). New York: Oxford University Press.

Hayek, F. A. (1960). *The Constitution of Liberty*. Chicago: University of Chicago Press.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker (2006). Costly punishment across human societies. *Science 312*, 1767–1770.

Hong, L. and S. E. Page (2001). Problem solving by heterogeneous agents. *Journal of Economic Theory 97*, 123–163.

Kornhauser, L. A. (2004). Governance structures, legal systems, and the concept of law. *Chicago-Kent Law Review 79*, 355–381.

Kramarz, F. (1996). Dynamic focal points in n-person coordination games. *Theory and Decision 40*, 277–313.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics 1*, 593–622.

Lewis, D. (2002 [1969]). *Convention*. Oxford: Harvard University Press.

Mahdi, N. Q. (1986). Pukhtunwali: Ostracism and honor among the pathan hill tribes. *Ethology and Sociobiology 7*, 295–304.

Mailath, G. J., S. Morris, and A. Postlewaite (2001). Laws and authority. Manuscript, University of Pennsylvania.

Mailath, G. J., S. Morris, and A. Postlewaite (2007). Maintaining authority. Manuscript, University of Pennsylvania.

Marmor, A. (1998). Legal conventionalism. *Legal Theory 4*, 509–531.

Marmor, A. (2009). *Social Conventions: From Language to Law*. Princeton: Princeton University Press.

McAdams, R. H. (2000). A focal point theory of expressive law. *Virginia Law Review 86*, 1649–1729.

McAdams, R. H. (2005). The expressive power of adjudication. *University of Illinois Law Review 2005*, 1043–1121.

Milgrom, P. R., D. C. North, and B. R. Weingast (1990). The role of institutions in the revival of trade: The medieval law merchant, private judges, and the champagne fairs. *Economics and Politics 2*, 1–23.

Myerson, R. B. (2004). Justice, institutions, and multiple equilibria. *Chicago Journal of International Law 5*, 91–107.

Ordeshoook, P. (1992). Constitutional stability. *Constitutional Political Economy 3*, 137–175.

Postema, G. J. (1982). Coordination and convention at the foundation of law. *The Journal of Legal Studies 11*, 165–203.

Raz, J. (1977). The rule of law and its virtue. *Law Quarterly Review 93*, 195.

Schelling, T. C. (1981). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sugden, R. (2005). *The Economics of Rights, Cooperation and Welfare* (2 ed.). New York: Palgrave Macmillan.

Sunstein, C. (1996). The expressive function of law. *University of Pennsylvania Law Review 144*, 2021–2053.

Weingast, B. R. (1997). The political foundations of democracy and the rule of law. *American Political Science Review 91*, 245–263.

Wiessner, P. (2005). Norm enforcement among the ju/'hoansi bushmen: A case of strong reciprocity? *Human Nature 16*, 115–145.

Notes

*We have benefitted from numerous conversations on an earlier draft of this paper. For detailed suggestions and references to related literature we are particularly grateful to Ken Arrow, Sam Bowles, Chuck Cameron, Tino Cuellar, Jean Ensminger, John Ferejohn, Les Green, Al Klevorick, Lewis Kornhauser, Antoine Lallour, Steve Macedo, Andrei Marmor, Richard McAdams, Tori McGeer, Mitch Polinsky, Dan Posner, Dan Ryan, Bill Simon, Jed Stiglitz, Joel Trachtman, Philip Pettit, and John Wallis.

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

[2]This focus on penalties does not deny that many systems we clearly would recognize as legal grant benefits–such as rights and entitlements. But any such benefit entails a correlative penalty—imposed on those who deny beneficiaries their rights or entitlements. An individual's right to a tax subsidy, for example, is implemented by classifying official behavior as wrongful if the subsidy is not paid.

[3]In this case, the classification might be better described as one that partitions punishable from not punishable behavior, although for expositional purposes, and with the proviso that we do not presume an inherent orientation to avoid wrongful conduct, we will use the term wrongful throughout.

[4]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 landlord and our buyers, serfs. Serfs then invest upfront in working the land, at the risk that in the future the landlord will extract a wrongful share of the harvest or wrongfully impose additional duties or conditions or hardships.

[5]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 lan-*

guage (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).

[6]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.

[7]"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)

[8]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.

[9]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.

[10]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 sub-game perfect equilibrium in the absence of formal and centralized legal penalties. Sub-game perfection in Greif's model of the Maghribi traders in the eleventh Century (as in a version of North, Milgom & 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 an agent who has cheated someone else in the past by refusing to hire him because

the cheater will, in equilibrium, cheat the new merchant as well.

[11]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 (2010) discusses the emerging role for private production of legal systems in globalized settings.