Anger and Enforcement

Robert J. Akerlof*

February 1, 2015

Abstract

Observers who are angered by rule violations and punish violators often play a critical role in enforcement. Hence a key question is: when will noncompliance provoke anger, and when will it be excused? This paper develops a theory of rule compliance as the outcome of a two-person Bayesian game. The core of the model is its description of what constitutes an excuse. Noncompliance is excused when a "reasonable person" in similar circumstances would also have failed to comply. Phenomena explained include the role of "legitimacy" in enforcement; corruption traps; graduated sanctions for repeat offenders; and tolerance of self-interestedness in markets.

Keywords: enforcement, rules, norms, legitimacy, anger.

JEL classification: D02, D03, K42, Z13.

*University of Warwick, r.akerlof@warwick.ac.uk. I would especially like to thank Robert Gibbons and Laurence Kotlikoff for their suggestions and support. I am also grateful to Philippe Aghion, George Akerlof, Oliver Hart, Hongyi Li, Rocco Macchiavello, Pascal Michaillat, Motty Perry, Philip Reny, Hui Tong, Jan Zilinsky, and seminar participants at ANU, Berkeley, Case Western, MIT, NBER, Toulouse, UNSW, The University of Sydney, The University of Warwick, and the Thurgau Experimental Economics Meeting.

1 Introduction

The enforcement of rules is central to the operation of all societies and organizations. Observers who are angered by rule violations and punish the violators often play a critical role in enforcement. That punishment may be as simple as reporting offenses to authorities. Enforcement may be difficult, however, if observers excuse noncompliance. The question arises: when will noncompliance be viewed as negligent, provoking anger and punishment, and when will it be excused?

In the law, the standard test of negligence is "the reasonable person test," which is defined as follows: "Negligence arises from doing an act that a reasonable person would not do under the circumstances, or from failing to do an act that a reasonable person would do."[1,2] This paper suggests that the reasonable person test is not only a principle of the law but, in fact, a basic feature of how people assess negligence. The paper builds a simple model in which agents use such a test to assign blame. It considers the implications for the enforcement of rules. In particular, we show that the model captures two important enforcement problems that are previously unaccounted for theoretically.

The model is a two-person, sequential-move game. The first player decides whether to comply with a rule, at a cost; the second player decides how much to punish him for any noncompliance, at a cost. We assume: (1) the first player could be of two types, one of which has a "sense of duty" to comply with the rule; and (2) the second player could be of two types, one of which holds that the first player has a duty to comply with the rule (or, put differently, defines a "reasonable" first player as one possessing a sense of duty). We critically assume that the second player is only angered by noncompliance if she thinks a "reasonable" first player would have complied in similar circumstances. Otherwise, noncompliance is excused.

In the model, anger over rule violations depends upon whether reasonable people are

[1] Miller and Perry (2012), p. 2.

[2] According to Unikel (1992): "From its modest beginnings, 'reasonableness' has gained a prominent position in almost every area of American law. A general survey reveals that the concept of 'reasonableness' is a standard of decision making in administrative law, bailment law, constitutional law, contract law, criminal law, tort law, and the law of trusts." (p. 327)

expected to comply; whether reasonable people comply depends, in turn, upon the anger over – and punishment of – violations. We consider the equilibria of this system. First, consider what equilibria look like when the second player feels there is a duty to comply, and defines a reasonable person accordingly. We find that, when the first player's cost of compliance is high, there is a unique equilibrium in which the first player fails to comply and the second player excuses noncompliance. When the first player's cost of compliance is low, there is a unique equilibrium in which the first player complies and the second player is angered by noncompliance.

When the cost of compliance takes an intermediate value, there are multiple equilibria. To illustrate, consider an example. Suppose there is a rule against corruption. There is the possibility of being in a good, low-corruption equilibrium. In this equilibrium, reasonable people are not corrupt; in consequence, corruption provokes anger and punishment; this threat of punishment sustains the low rate of corruption. But, there is also the possibility of being trapped in a high-corruption equilibrium. In this equilibrium, corruption is rampant, with even reasonable people (with a sense of duty) engaging in it; in consequence, it is excused; the lack of punishment sustains the high rate of corruption.

The possibility of a corruption trap, in which the endemic nature of corruption leads people to excuse it, is the first important enforcement problem identified by the paper. This result is consistent with a large empirical literature on corruption (see especially Collier (2000) and Rose-Ackerman (2001)). There are, in fact, a number of other "noncompliance traps" that have been documented.

The paper identifies a second enforcement problem that is, perhaps, even more fundamental. If the second player feels there is no duty to comply – and defines a reasonable person accordingly – there is a unique equilibrium with the property that noncompliance is always excused. It follows that absence of sense of duty to follow rules presents a serious challenge to enforcement. If the first player lacks a sense of duty to comply, he complies only to avoid punishment; if the second player feels there is no duty to comply, she excuses noncompliance and hence inflicts no punishment.

This finding relates to a large body of work outside of economics, on "legitimacy." By common definition, a rule is "legitimate" if there is a widespread feeling that there is a duty to comply. Complete lack of legitimacy thus corresponds to the case in the model where neither player feels there is a duty to comply. Blau (1964) argues that rules will be disobeyed in the absence of legitimacy because "coercive use of power engenders resistance."[3] That resistance, according to Ostrom (1990), is commonly manifested in reluctance to report violations to authorities. Violations are not reported because they are excused. In consequence, "the legitimacy of rules... will reduce the costs of monitoring, and [its] absence will increase [the] costs."[4] Later in the paper, we will consider a number of examples where enforcement problems arise from lack of legitimacy. They suggest that the need to legitimate rules serves as a constraint; this constraint often has large effects on the way in which organizations are structured.

The model yields a theory of anger as well as a theory of rule enforceability. It makes several predictions regarding anger and punishment. First, as we have already indicated in passing, anger depends upon the cost of compliance. A high cost of compliance serves as an excuse for breaking a rule. This result is intuitive. It is also consistent with survey evidence from Kahneman, Knetsch, and Thaler (1986). They found, for example, that 68 percent of survey respondents excused a company for reducing wages when it was losing money; and 75 percent excused a landlord for raising the rent on a property when the landlord's costs increased, even though increasing the rent meant the current tenant would be forced to move.[5]

Provided the second player does not know the first player's cost of compliance, we find that anger-over-noncompliance depends upon beliefs about the first player's type. The logic is as follows. Upon observing noncompliance, player 2 is uncertain whether there is

[3] Blau (1964), pp. 199-200.

[4] Ostrom (1990), p. 204.

[5] A high cost of compliance is similarly seen as a valid excuse in courts of law, for which there are many legal precedents. The case of United States v. Carroll Towing Company is an important precedent in the use of the reasonable person test. The case concerned a barge that broke adrift, collided with a tanker, and sank. In assessing whether there was negligence on the part of the barge owner, Judge Learned Hand tried to determine what a reasonable person would have done. He decided, in particular, that what a reasonable person would be expected to do depended upon the cost of taking precautions.

a good excuse (the cost of compliance is high) or player 1 is just unreasonable. Player 2's prior on player 1's type affects the weight she puts on there being a good excuse. This result explains the empirical finding that anger escalates with repeated noncompliance (since repeated noncompliance is likely to signal that a person lacks a sense of duty).[6]

Finally, it has been observed that anger is contextual. For instance, there is typically a high degree of tolerance of self-interested behavior in markets. Sen (1977) has seen this as a puzzle, given people's strong concerns for fairness in other settings. The model accounts for the high tolerance of self-interested behavior in markets. This is explained as being due to context-specific definitions of what is reasonable.

Relation to Existing Literature. First and foremost, the paper contributes to the economic literature on norms (for a survey of approaches to norms, see Young (2008)). Norms are often conceptualized as internalized views regarding duty and obligation (see, for example, Elster (1989), Besley and Ghatak (2005), Akerlof and Kranton (2005), Prendergast (2007, 2008), and Benabou and Tirole (2003)).[7] In stark contrast, norms are not internalized at all in other work. Agents simply obey norms in order to avoid punishment (see, for instance, Kandori (1992)).[8] A question is whether there is, in fact, a relationship between punishment and internalized conceptions of duty. A number of scholars have posited that there is (see, for instance, Sugden (1986) and Coleman (1990)). There is also considerable experimental work demonstrating people's willingness to inflict costly punishment when internalized norms are violated.[9]

This paper explains how views regarding duty can give rise to sanctions: a player's feeling that there is a duty to comply not only motivates compliance, it potentially motivates punishment of others' noncompliance. Norm violations sometimes provoke anger and punishment in the model, but, on the other hand, they may be excused. The paper further explains when noncompliance will and will not be punished. Moreover, it describes when

[6] See the discussion of Kliemann, Young, Scholz, and Saxe (2008) in Section 5.

[7] Some of these papers use different terminology, such as "intrinsic motivation" or "values."

[8] Young (2008) lists two additional reasons why people may comply with norms. They may do so in order to coordinate with others (see Young (1993,1996)). They may also obey in order to avoid social stigma (see Bernheim (1994), Lindbeck, Nyberg, and Weibull (1999), and Benabou and Tirole (2011)).

[9] See, for example, Fehr, Fischbacher, and Gächter (2002)).

those who lack a sense of duty (have failed to internalize the norm) may nonetheless comply in order to avoid such sanctions.

The paper also contributes to the literature on rule enforcement. The model, in particular, is unique in identifying legitimacy of rules as a determinant of enforceability. Standard approaches to rule enforcement are repeated-game models and third-party enforcement models (see especially Becker (1968)), neither of which describe the role of legitimacy. Legitimacy is highly stressed outside of economics – in particular, in the legal literature (see, for instance, our later discussion of Tyler (1990) and Fagan and Meares (2008)). A point that is particularly stressed is that when punishments are too harsh – and fail to fit the crimes – it makes enforcement more difficult. Such punishments may undermine the legitimacy of rules/laws. A recent economics paper, Chen (2013), provides some empirical confirmation. Chen (2013) looks at a dataset of British and Irish soldiers sentenced to death by the British military during World War I. He finds that executions of British deserters deterred absences; in contrast, executions of non-deserters and Irish soldiers spurred absences. Another related paper is Benabou and Tirole (2011). They consider how the law affects agents' internalized values – as well as the social stigma associated with crime.

There is also a literature related to the other enforcement problem discussed in the paper – corruption traps (see Bardhan (1997) for a review). Our paper provides a novel and important account of corruption traps. In particular, it is the only model in which a high rate of corruption leads agents to excuse it. However, there are other stories – complementary to our own –such as: the difficulty of auditing corrupt officials in corrupt societies (Lui (1986), Cadot (1987), and Andvig and Moene (1990)); the higher returns to corruption relative to entrepreneurship when corruption is more prevalent (Murphy, Shleifer, and Vishny (1991, 1993), and Acemoglu (1995)); and the greater willingness of individuals to engage in corruption when others are believed to be doing so (Sah (1988)).

As mentioned earlier, in addition to yielding a theory of rule enforceability, the model also yields a theory of anger. It makes predictions regarding when one player will blame – or feel unfairly treated by – another. It thus relates to work on fairness. In contrast to existing

fairness models, in which there is a fixed notion of what constitutes fair treatment, what is considered fair here varies with the definition of what is reasonable. This allows the model to account for contextual differences in fairness attitudes, such as the particularly high tolerance of self-interested behavior in markets. There is, however, overlap with existing models. In their survey article, Fehr and Schmidt (2006) distinguish two types of reciprocity models of fairness: "intention-based models" (such as Rabin (1993)) and "interdependent-preference models" (such as Levine (1998)).[10] Our model is a hybrid, and it addresses respective concerns: for example, in intention-based models, a high cost of providing a benefit to others does not excuse the failure to do so; in interdependent-preference models, those known to be selfish are punished even when they behave just like unselfish types.[11,12]

Finally, the paper makes a purely technical contribution, which is worth highlighting. While the equilibrium concept applied in our analysis is a standard one (Perfect Bayesian Equilibrium), the game tree (see Figure 2, page 13) is not. In contrast to standard games, the particular end node of the game tree players reach does not fully determine their payoffs. Payoffs, at an end node, still depend upon players' choice of strategies. The reason is that the second player's utility depends upon the answer to a potentially counterfactual question: what would a reasonable person have done under similar circumstances? The idea that counterfactual scenarios can matter for payoffs may have applicability beyond the present model.

The paper proceeds as follows. Section 2 gives a simple example. Section 3 describes the formal model. Section 4 solves for the equilibria of the game. Section 5 discusses four

[10] In "intention-based models," players are of a single type; they treat opponents kindly (hostilely) when they believe opponents intend to treat them kindly (hostilely). In "interdependent-preference models," there are two types: selfish players are always selfish, while altruistic players are kind when they believe their opponents are altruistic and selfish when they believe their opponents are selfish.

[11] In interdependent-preference models, those known to be selfish are punished even when they behave the same way as others because punishment is meted out solely by type. In contrast, in the model developed in this paper, unreasonable people may avoid punishment by behaving as a reasonable person would behave.

[12] Our model also addresses concerns with a third category of fairness model, in which players have "social preferences." In social preference models, players care about the material allocations both to themselves and to others. The theory of fairness described in this paper has some potential overlap: for example, players might consider it a duty to show concern for inequity. But "duty" is defined much more generally. Our model can also account for the many situations in which people tolerate self-interested behavior, since it need not be a duty to show concern for others.

phenomena captured by the model: the role of "legitimacy" in enforcement, noncompliance traps, graduated sanctions for repeat offenders, and tolerance of self-interestedness in markets. Section 6 concludes.

2 An Illustrative Example

The key aspects of the theory can be illustrated with a simple example. Consider a game with two players: 1 and 2. Player 1 is of type $I_1 \in \{0, 1\}$. $I_1 = 1$ with probability q and $I_1 = 0$ with probability $1-q$, with $q < 1$. An $I_1 = 0$ type has instrumental preferences, whereas an $I_1 = 1$ type feels some sense of duty to comply with a rule. Player 1 knows his type but player 2 does not know player 1's type.

At time 1, player 1 chooses whether to comply with the rule ($a = 1$), or not ($a = 0$). The cost of complying is $C > 0$. An $I_1 = 1$ type also faces a cost $D > 0$ if he fails to comply, reflecting his sense of duty to do so.[13] Compliance is valued by player 2 in amount v.

In the event of noncompliance, player 2 decides at time 2 how much to punish player 1: $p \geq 0$. We assume player 2's preferences are such that she chooses $p = \theta M$ (for a more detailed discussion of player 2's preferences, see the next section). θ denotes player 2's ability to punish. M denotes player 2's feeling of mistreatment. M is defined as how much worse off player 2 is than she would be had player 1 been a "reasonable person." We assume player 2 defines a reasonable person as someone with a sense of duty to comply (an $I_1 = 1$ type).[14] Let a^{rp} denote the compliance choice of a reasonable ($I_1 = 1$) type. If a reasonable person would have complied ($a^{rp} = 1$) but player 1 fails to comply ($a = 0$), $M = v$. Otherwise, $M = 0$.

The equilibrium concept we will apply is formally described in Section 3. For the purposes of this example, we can think of an equilibrium as a pair (a^{rp^*}, p^*) satisfying

[13] While we assume for the purpose of simplicity that player 1 loses D regardless of the cost of compliance C, the model can easily be generalized so that D is a function of C. This accounts for circumstances in which a dutiful type only feels a duty to comply when it is not too costly to do so.

[14] In the formal model, we will also consider what happens when player 2 considers it reasonable to be an $I_1 = 0$ type.

the following two conditions: (1) taking the punishment of noncompliance p^* as given, the reasonable type ($I_1 = 1$) finds it optimal to choose $a = a^{rp^*}$; and (2) taking the behavior of a reasonable person (a^{rp^*}) as given, player 2 finds it optimal to inflict punishment p^* in the event of noncompliance.[15]

Condition (1) can be restated as: $a^{rp^*} = 1$ if $C \leq p^* + D$ and $a^{rp^*} = 0$ otherwise.[16] Condition (2) can be restated as: $p^* = \theta v a^{rp^*}$. It follows from these conditions that two types of equilibria can arise. An equilibrium with punishment ($a^{rp^*} = 1, p^* = \theta v$) exists if $C \leq \theta v + D$. An equilibrium with no punishment of noncompliance ($a^{rp^*} = 0, p^* = 0$) exists if $C > D$.

In the equilibrium with punishment, the $I_1 = 1$ type complies and, if $C \leq \theta v$, the $I_1 = 0$ type also complies. The $I_1 = 0$ type, who is instrumental, complies purely in order to avoid punishment θv. In the equilibrium with no punishment, neither the $I_1 = 1$ type nor the $I_1 = 0$ type complies.

Observe that the types of equilibria that arise depend upon whether the cost of compliance is low ($C \leq D$), intermediate ($D < C \leq D + \theta v$), or high ($C > D + \theta v$). When the cost of compliance is relatively low ($C \leq D$), there is a unique equilibrium in which noncompliance is punished ($p^* = \theta v$). Reasonable people comply even if noncompliance is not punished, so player 2 is necessarily angered by noncompliance.

When the cost of compliance is relatively high ($C > \theta v + D$), there is a unique equilibrium in which noncompliance is not punished ($p^* = 0$). The cost of compliance is sufficiently high that a reasonable person cannot be induced to comply. Player 2 excuses noncompliance since reasonable people do not comply.

[15] Condition (1) corresponds to the standard rationality assumption for player 1 in a Bayesian game. Condition (2) encompasses assumptions about both the on- and off-equilibrium-path behavior of player 2.

Our focus in Section 3 will be on perfect Bayesian equilibria (PBE). Condition (2) corresponds to a refinement of PBE, which we refer to as Reasonable-Person (RP) Stability. See the Appendix for a discussion.

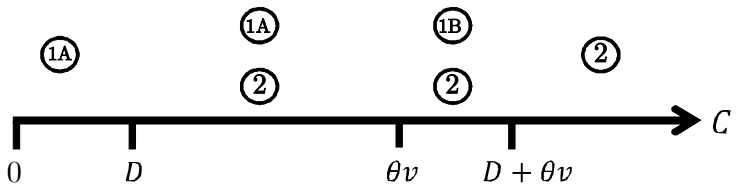
The set of PBE and the set of RP-stable PBE look very similar. They generate exactly the same compliance behavior but PBE admits a larger set of possible punishments of noncompliance.

In the present example, there is a unique RP- stable PBE of type 1A (see Figure 1), in which both the $I_1 = 0$ and $I_1 = 1$ types comply; such an equilibrium exists when $C \leq \theta v$ and noncompliance receives punishment $p^* = \theta v$. There are, in contrast, a set of type 1A PBE; such equilibria exist when $C \leq \theta v$ and noncompliance receives punishment $p^* \in [\min(\theta v, C), \theta v]$. This is the only difference.

[16] We are assuming here, for simplicity, that player 1 complies if he is otherwise indifferent.

When the cost of compliance takes an intermediate value ($D < C \leq \theta v + D$), there are multiple equilibria. In one type of equilibrium, there is no compliance and no punishment of noncompliance. In the other type of equilibrium, there is compliance (by the $I_1 = 1$ type only or by both types) and punishment of noncompliance. The reason for multiple equilibria is as follows. In the first type of equilibrium, reasonable people fail to comply ($a^{rp} = 0$). As a result, player 2 excuses noncompliance and fails to punish it ($p = 0$). The lack of punishment makes it optimal for reasonable people to choose noncompliance. In the second type of equilibrium, reasonable people comply ($a^{rp} = 1$). As a result, there is considerable anger over (and punishment of) noncompliance ($p = \theta v$). This punishment is sufficient to induce compliance from reasonable people.

Figure 1 illustrates the equilibria for the case where $D < \theta v$.



1A: punishment θv of noncompliance, compliance by both $I_1 = 0,1$ types.
1B: punishment θv of noncompliance, compliance by the $I_1 = 1$ type only.
2: no punishment, no compliance.

Figure 1: Equilibria of the Game

This example demonstrates important features of the theory. First, anger over noncompliance depends upon the cost of compliance. Noncompliance is excused when the cost of compliance is high. There are multiple equilibria when the cost of compliance takes an intermediate value. Second, we see that player 2's definition of reasonableness – whether player 2 feels there is a duty to comply (as we have assumed here) or not – matters for what will be excused.

In this example, player 2's anger over (and punishment of) noncompliance did not depend upon her prior regarding player 1's type (q). In contrast, player 2's prior does play

a role when player 1 does not know player 2's cost of compliance (C). Upon observing noncompliance, player 2 may be uncertain whether player 1 had a good excuse (C was high) or was just unreasonable. Player 2's prior about player 1's type (q) will affect how much weight player 2 puts on player 1 being unreasonable.

3 The Model

3.1 Statement of the Problem

We now turn to the formal model. Consider a two-period Bayesian game with two players: 1 and 2. Player 1 is of type $I_1 \in \{0, 1\}$, where $I_1 = 0$ reflects a player with instrumental preferences and $I_1 = 1$ reflects a player with a sense of duty to follow a rule.

Player 1 knows his own type (I_1) and the cost of compliance (C). Player 2 does not know I_1 or C. Player 2's prior on I_1 is that $I_1 = 1$ with probability q and $I_1 = 0$ with probability $1 - q$, with $q < 1$. Player 2's prior on C is that it is drawn from a distribution with cdf F and support $S \subseteq (0, \infty)$. I_1 and C are assumed to be independent. Let $\mu(I_1, C)$ denote player 2's joint prior on I_1 and C.

Player 2 also has a type $I_2 \in \{0, 1\}$. For simplicity, we assume I_2 is common knowledge.[17] If $I_2 = 1$, player 2 feels there is a duty to comply. If $I_2 = 0$, player 2 feels there is no duty (player 2 feels player 1 is entitled to do what he wants). I_2 describes how player 2 defines a reasonable person. Player 2 defines a reasonable person as an $I_1 = I_2$ type.[18]

At time 1, player 1 chooses whether to comply with a rule ($a \in \{0, 1\}$). We will refer to $a = 1$ as "compliance" and $a = 0$ as "noncompliance." Player 2 observes player 1's choice. If player 1 fails to comply, player 2 chooses how much to punish player 1 at time 2: $p \geq 0$.[19]

We restrict attention to pure strategies for the players. A strategy for player 1 is a

[17] See Section 3.2 for a discussion of this assumption. A natural alternative would be to assume player 1 does not know player 2's type (I_2) and believes $I_2 = 1$ with probability q.

[18] We might have chosen, alternatively, to describe a person as "reasonable" whenever $I_1 \geq I_2$. While we define a reasonable person as an $I_1 = I_2$ type, we will see that the model implies that, whenever player 2 excuses the behavior of an $I_1 = I_2$ type, she will also excuse the behavior of an $I_1 > I_2$ type.

[19] An alternative game where player 2 can punish player 1 after $a = 1$ would require a more elaborate equilibrium concept such as sequential equilibrium but would yield qualitatively similar results.

function $\tilde{a}(C, I_1)$. $\tilde{a}(C, I_1)$ denotes player 1's choice of a given C and I_1. A strategy for player 2 is $\tilde{p}$, which denotes player 2's choice of how much to punish noncompliance. (While player 2 also has a type I_2, I_2 can simply be treated as an exogenous parameter of the model. So, we choose not to write player 2's strategy as a function of her type.)

If player 1 chooses action a and player 2 follows strategy $\tilde{p}$, player 1's utility is given by:

$$U_1(a, \tilde{p}, C, I_1) = -\tilde{p} \cdot (1 - a) - C \cdot a - D \cdot \max(I_1 - a, 0).$$

The first term is the punishment player 2 inflicts on player 1. The second term is the cost incurred if player 1 complies. The third term is the loss of utility to player 1 from failing to do her duty. It is equal to zero unless player 1 feels he has a duty to comply and fails to do so ($I_1 = 1$ and $a = 0$), in which case it is equal to $D > 0$.[20] The first two terms of the utility function are the economic part of player 1's utility function, reflecting what player 1 *wants* to do, while the third term reflects what player 1 feels he *should* do.

Before defining player 2's utility function, it is first necessary to define mistreatment (M). In line with our previous discussion, we define mistreatment as how much better off player 2 would be if player 1 had been a reasonable person (had possessed an appropriate sense of duty). More formally, when player 1 plays strategy $\tilde{a}$:

$$M(\tilde{a}, C, I_1) = \max\left(v \cdot \tilde{a}^{rp}(C) - v \cdot \tilde{a}(C, I_1), 0\right).$$

$v \geq 0$ is how much player 2 values compliance.[21] $\tilde{a}^{rp}(C)$ denotes the strategy followed by a reasonable person. Since a reasonable person is defined as an $I_1 = I_2$ type, $\tilde{a}^{rp}(C) = \tilde{a}(C, I_2)$. $v \cdot \tilde{a}^{rp}(C)$ is what player 2 would receive from player 1 if player 1 were reasonable.

[20] As mentioned earlier, we assume player 1 loses D regardless of the cost of compliance C for simplicity. The model can be generalized so that D is a function of C in order to account for circumstances in which a dutiful type only feels a duty to comply when it is not too costly to do so.

The duty term could also be generalized in other ways. We can capture more sophisticated rules – such as, for example, a duty to adhere to orders given by a leader – by allowing duty to depend upon the state of the world (in this case, the order given by the leader).

[21] We remark, more than parenthetically, that the model is unaltered if player 2 herself is not harmed by player 1's noncompliance, but she is instead altruistic and loses altruistic utility v when player 1 fails to comply and this harms some third party. This broadens the interpretation of the model.

$v \cdot \tilde{a}(C, I_1)$ is what player 2 actually receives from player 1.[22] According to this formula, player 2 only feels mistreated when a reasonable person would comply ($\tilde{a}^{rp}(C) = 1$) but player 1 fails to comply ($\tilde{a}(C, I_1) = 0$), in which case $M = v$. Otherwise, $M = 0$.

If player 2 chooses to punish noncompliance in amount p, her utility is given by:

$$U_2(a, \tilde{a}, p, C, I_1) = v \cdot a - \kappa(p) \cdot (1 - a) - \Phi(M(\tilde{a}, C, I_1), p \cdot (1 - a)).$$

The first term is the benefit v player 2 receives when player 1 complies. The second term is the cost of inflicting punishment on player 1, where $\kappa(p)$ is the cost of inflicting punishment p. The final term motivates player 2 to punish player 1 when she feels mistreated ($M > 0$). Φ represents the disutility associated with feeling mistreated; this disutility is reduced by punishing player 1.

For the remainder of the paper, we will use the following functional forms for κ and Φ:[23]

$$\kappa(p) = \frac{p}{\theta},$$

$$\Phi(M, p) = M \log\left(\frac{kM}{p}\right).$$

$\theta > 0$ parameterizes the ability to punish player 1 (higher θ implies a greater ability to punish). Observe that the presence of a higher authority to which rule violations can be reported would serve to increase θ. In order to ensure that player 2's utility is decreasing in the mistreatment she suffers, we assume $k > \theta e^{-1}$.

Figure 2 illustrates the game tree for the case where C takes a single value. Observe that player 2's payoff at each node depends upon player 1's choice of strategy ($\tilde{a}$). While we typically do not see Bayesian games with this property, it is still a standard Bayesian

[22] As mentioned in the introduction, this definition of mistreatment exactly corresponds to the "reasonable person test," which is widely used in contract law, criminal law, and tort law (among other branches) to assess liability or guilt.

[23] These functional forms are convenient because they make it optimal for player 2 to choose $p = \theta M$ (or $p = \theta \cdot E(M)$ when M is unknown). The choice of functional forms is not critical however. What matters is that the functional forms for κ and Φ ensure that: player 2 punishes player 1 more when M is greater and when player 2's cost of punishment is lower.

$M \log\left(\frac{kM}{p}\right)$ is technically undefined when $M = 0$. We will assume $\Phi(0, p) = \lim_{M \to 0^+} M \log\left(\frac{kM}{p}\right) = 0$.

game and standard equilibrium concepts can be applied.

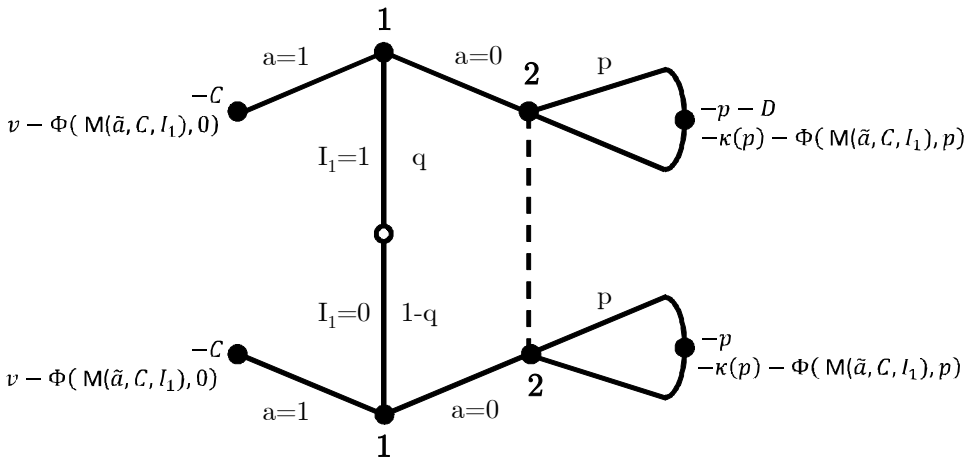


Figure 2: The game tree for the case where C takes a single value.

We will focus on perfect Bayesian equilibria of the game. For the sake of simplifying analysis, we assume in the equilibrium definition that player 1 chooses to comply if he is otherwise indifferent. Equilibria in the stylized version of the model presented in Section 2 correspond to a refinement of perfect Bayesian equilibrium. For a detailed discussion, see the Appendix.

The equilibrium definition, stated below, has three conditions. Condition D1 says that player 1 maximizes utility, taking into account player 2's punishment strategy (choosing compliance if indifferent). Condition D2 says that player 2 maximizes expected utility given her posterior beliefs $\mu(\cdot|a)$. Condition D3 says that player 2 updates her beliefs according to Bayes' rule whenever possible. If player 1 chooses a with zero probability, so that Bayes' rule is undefined, any posterior beliefs are then admissible.

Definition 1 *A strategy pair* $(\tilde{a}^*, \tilde{p}^*)$ *is a perfect Bayesian equilibrium if:*

(D1) $\tilde{a}^*(C, I_1) \in \arg\max_a U_1(a, \tilde{p}^*, C, I_1)$ *for all* C, I_1.

$\tilde{a}^*(C, I_1) = 1$ *if* $1 \in \arg\max_a U_1(a, \tilde{p}^*, C, I_1)$ *for all* C, I_1.

(D2) $\tilde{p}^* \in \arg\max_p E_{\mu(\cdot|0)}[U_2(0, \tilde{a}^*, p, C, I_1)]$.

(D3) $$\mu(C, I_1|a) = \begin{cases} \frac{\mu(C,I_1)}{\mu(\{(C',I_1'):\tilde{a}^*(C',I_1')=a\})}, & \text{if } \tilde{a}^*(C, I_1) = a \\ 0, & \text{otherwise} \end{cases}$$
$$\text{if } \mu(\{(C', I_1') : \tilde{a}^*(C', I_1') = a\}) > 0.$$

3.2 Discussion of the Model

Let us pause briefly to discuss a few modeling choices.

First, we assumed player 1's type is drawn from a distribution ($I_1 = 1$ with probability q and $I_1 = 0$ otherwise); in contrast, we fixed player 2's type (I_2) and assumed it was common knowledge. This leads to the possibility player 1 lacks a sense of duty with high probability (q low) but still expects to face a second player who feels there is a duty to comply ($I_2 = 1$). A natural alternative, which deals with this issue, would be to assume player 1 does not know player 2's type and holds the prior that $I_2 = 1$ with probability q. In this case, both players' types are drawn from the same distribution. We chose not to adopt this assumption in order to simplify the model and analysis. Given that the reasonable person test is a new idea, we consider it best to present the simplest possible formulation. Furthermore, this alternative assumption does not yield much additional insight.

Second, the model suggests various ways in which q (the probability of a dutiful type) might be endogenized. For instance, one could imagine a many-round version of the game, in which multiple pairs of agents play each round. The prevalence of dutiful types in round $t + 1$ (q_{t+1}) could depend upon the rate of compliance in round t. This would capture the idea that sense-of-duty might erode if rule-breaking were pervasive. We chose not to endogenize q because it would be a distraction from our focus, which is to understand the implications of the reasonable person test. Nonetheless, a model in which q were endogenized would be a natural and worthwhile extension.

Third, the assumptions we made regarding player 2's utility function may seem strong. However, these assumptions are simply meant to rationalize player 2 punishing player 1 more when: (1) she feels more mistreated (M is larger), and (2) she has a greater ability to punish (θ is larger).

Fourth, player 1 also has a conception of what is reasonable; it would be natural to put into player 1's utility function the same machinery as exists in player 2's, so that he might feel angered. If we further gave player 1 the opportunity to punish player 2, this would lead to the possibility of retaliatory punishment. Such a model would be a natural extension; it presents the possibility of capturing feuding. Again, for the sake of simplicity, we have ruled out the possibility of retaliation.

Fifth, we assumed player 1 loses D from failing to comply regardless of the cost of compliance. This assumption was also made for the sake of simplicity. We could make the model more realistic by assuming D is a function of the cost of compliance, so that a dutiful type does not feel it is a duty to comply when it is too costly to do so.

Finally, it is worth highlighting that, even though the equilibrium concept we will use is standard (PBE), the game tree is not. Normally, the end node players reach is sufficient to determine their payoffs. In our game, at an end node, player 2's payoff depends upon player 1's strategy. This feature of the game is due to player 2's use of the reasonable person test.

3.3 Characterizing the Equilibria

Our assumptions on player 1's utility allow us to restate condition D1 of definition 1 as follows:

$$\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p}^* + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$$

If player 1 has instrumental preferences ($I_1 = 0$), player 1 complies when the punishment of noncompliance is greater than or equal to the cost of compliance: $C \leq \tilde{p}^*$. Player 1 is more compliant if he feels a sense of duty to comply: player 1 complies if $C \leq \tilde{p}^* + D$.

Our assumptions on player 2's utility function allow us to restate condition D2 of definition 1 as follows:

$$\tilde{p}^* = \theta \cdot E_{\mu(\cdot|0)}[M(\tilde{a}^*, C, I_1)]. \qquad (*)$$

Player 2's punishment of noncompliance is increasing in her ability to punish (θ) and her expectation of how much she has been mistreated (M).

First, consider the case where player 2 feels there is no duty to comply ($I_2 = 0$). When $I_2 = 0$, $M(\tilde{a}^*, C, I_1) = 0$ for all values of C and I_1.[24] Intuitively, player 2 cannot feel mistreated when $I_2 = 0$ since player 1 is always at least as compliant as a reasonable person. Hence, condition (*) implies that noncompliance is excused ($\tilde{p}^* = 0$) when player 2 feels there is no duty to comply ($I_2 = 0$).

Now, consider the case where player 2 feels there is a duty to comply ($I_2 = 1$). When Bayes' rule is applicable, condition (*) implies that:

$$\tilde{p}^* = \theta v \cdot \frac{[F(\tilde{p}^* + D) - F(\tilde{p}^*)](1-q)}{[1 - F(\tilde{p}^*)] - q[F(\tilde{p}^* + D) - F(\tilde{p}^*)]}.$$

where F is player 2's prior on the distribution of C and q is player 2's prior on the probability that player 1 is of type $I_1 = 1$. It can be shown that Bayes' rule is applicable whenever $F(\tilde{p}^*) < 1$.[25]

When $F(\tilde{p}^*) = 1$, Bayes' rule is not applicable and condition D3 places no restrictions on posterior beliefs ($\mu(\cdot|0)$). In this case, player 2's expectation of how much she has been mistreated ($E_{\mu(\cdot|0)}[M(\tilde{a}^*, C, I_1)]$) can take any value in the interval $[0, v]$. Hence, condition (*) only requires that $\tilde{p}^* \in [0, \theta v]$.[26]

Lemma 1 summarizes, giving necessary and sufficient conditions for a pair of strategies to be a perfect Bayesian equilibrium. Formal proofs are given in the Appendix.

Lemma 1 *A pair of strategies $(\tilde{a}^*, \tilde{p}^*)$ is a perfect Bayesian equilibrium if and only if:*

(P1) $$\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p}^* + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$$

(P2) *If $F(\tilde{p}^*) < 1$:*

[24] The mathematical reasoning is as follows. Recall that $M(\tilde{a}^*, C, I_1) = v \cdot \max(\tilde{a}^*(C, I_2) - \tilde{a}^*(C, I_1), 0)$. From our restatement of condition D1, we know that $\tilde{a}^*(C, I_1)$ is nondecreasing in I_1. Hence, $\max(\tilde{a}^*(C, I_2) - \tilde{a}^*(C, I_1), 0) = 0$ whenever $I_2 = 0$.

[25] See the proof of Lemma 1 in the Appendix for a derivation of these results.

[26] The proof of Lemma 1 also derives this result.

$$\tilde{p}^* = \begin{cases} 0, \text{ if } I_2 = 0 \\ \theta v \cdot \frac{[F(\tilde{p}^*+D)-F(\tilde{p}^*)](1-q)}{[1-F(\tilde{p}^*)]-q[F(\tilde{p}^*+D)-F(\tilde{p}^*)]}, \text{ if } I_2 = 1 \end{cases}$$

(P3) *If* $F(\tilde{p}^*) = 1$*:*

$\tilde{p}^* = 0$, *if* $I_2 = 0$

$\tilde{p}^* \in [0, \theta v]$, *if* $I_2 = 1$

4 Results

We will now consider the equilibria of the game: first, describing the case where player 2 feels there is no duty to comply ($I_2 = 0$); then, turning to the case where player 2 feels there is a duty ($I_2 = 1$).

How much will player 2 punish noncompliance when player 2 feels there is no duty to comply ($I_2 = 0$)? It follows from Lemma 1 that there is a unique perfect Bayesian equilibrium. Noncompliance is not punished ($\tilde{p}^* = 0$) and player 1 complies only from a sense of duty (only if $I_1 = 1$ and $C \leq D$).

Proposition 1 *If player 2 feels there is no duty to comply* ($I_2 = 0$)*, a unique perfect Bayesian equilibrium exists. In this equilibrium, player 2 does not punish noncompliance:*

$$\tilde{p}^* = 0$$

$$\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$$

What drives the finding that noncompliance is excused? Player 2 feels mistreated when she believes she has been harmed because player 1 lacks an appropriate sense of duty ($I_1 < I_2$). In this case, either player 1 has a minimally appropriate sense of duty ($I_1 = I_2 = 0$) or player 1 has a greater sense of duty to comply than player 2 thinks is required ($I_1 = 1 > I_2 = 0$). Since noncompliance cannot be due to a lack of an appropriate sense of duty, player 2 will not feel mistreated when player 1 fails to comply.

Proposition 1 may seem trivial, but it has important implications. It says that, if player 2 feels player 1 has no duty to comply (is entitled to do what he wants), player 2 will not get

angry when player 1 behaves in a self-interested way. It gives a rationale for self-interested behavior sometimes being acceptable and not provoking anger.

We turn now to the case where player 2 feels there is a duty to comply ($I_2 = 1$). Proposition 2, stated below, considers the case in which player 2 believes $C = C_{low}$ with probability r and $C = C_{high} \geq C_{low}$ with probability $1 - r$. The special case in which $C_{low} = C_{high} = \bar{C}$ corresponds to the illustrative example from Section 2.

Proposition 2 *Suppose player 2 feels there is a duty to comply ($I_2 = 1$), and player 2 believes $C = C_{low} > 0$ with probability r and $C = C_{high} \geq C_{low}$ with probability $1 - r$ $(0 < r < 1)$. There are three types of perfect Bayesian equilibria that can arise.*[27]

Type 1: *the reasonable type ($I_1 = 1$) always complies.*

If $C_{high} \leq D + \theta v$, PBE exist with:

(i) $\tilde{p}^* \in [\min(\theta v, C_{high}), \theta v]$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p}^* + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

Type 2: *the reasonable type never complies.*

If $C_{low} > D$, a PBE exists with:

(i) $\tilde{p}^* = 0$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

Type 3: *the reasonable type sometimes complies (when $C = C_{low}$).*

If $\left(\frac{r-rq}{1-rq}\right)\theta v < C_{low} \leq D + \left(\frac{r-rq}{1-rq}\right)\theta v < C_{high}$, a PBE exists with:

(i) $\tilde{p}^* = \left(\frac{r-rq}{1-rq}\right)\theta v$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \left(\frac{r-rq}{1-rq}\right)\theta v + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

Three types of equilibria can arise. The behavior of the reasonable type ($I_1 = 1$) differs across these equilibria. In an equilibrium of type (1), reasonable people always comply.

[27] If $C_{low} \leq \min\left(D, \left(\frac{r-rq}{1-rq}\right)\theta v\right)$ and $C_{high} > D + \theta v$, an equilibrium does not exist. This is consistent with Lemma 2 (see below), since Lemma 2 does not ensure existence in this particular case.

In an equilibrium of type (2), reasonable people never comply. In an equilibrium of type (3), reasonable people sometimes comply (when $C = C_{low}$). Noncompliance is excused in a type (2) equilibrium since reasonable people do not comply. Noncompliance is punished in type (1) and type (3) equilibria since reasonable people do sometimes comply.

Let us begin by discussing the case where $C_{low} = C_{high} = \bar{C}$, corresponding to the illustrative example. Only equilibria of types (1) and (2) arise when $C_{low} = C_{high} = \bar{C}$. The set of perfect Bayesian equilibria produce a picture nearly identical to Figure 1. The only difference is that a wider range of punishments are admissible in Region 1A: $\tilde{p}^* \in [\min(\theta v, C_{high}), \theta v]$ rather than $\tilde{p}^* = \theta v$. These equilibria do not survive an appropriate refinement of PBE (see Appendix).

As in the illustrative example, the set of perfect Bayesian equilibria depends upon whether the cost of compliance is low, intermediate, or high. Let us recall the reason for this result. When the cost of compliance is low ($C \leq D$), reasonable people necessarily comply, so noncompliance is not excused. When the cost of compliance is relatively high ($C > \theta v + D$), reasonable people cannot be induced to comply. So, noncompliance is always excused.

When the cost of compliance takes an intermediate value ($D < C \leq \theta v + D$), there are equilibria of both types. The reason for multiple types of equilibria is as follows. When reasonable people comply, there is considerable anger over (and punishment of) noncompliance. This punishment is sufficient to induce reasonable people to comply. If reasonable people fail to comply, however, player 2 excuses noncompliance and does not punish it. The lack of punishment makes it optimal for reasonable people to choose noncompliance.

When $C_{high} > C_{low}$, equilibria of type (3) sometimes exist. In an equilibrium of type (3), noncompliance receives punishment of $\tilde{p}^* = \left(\frac{r - rq}{1 - rq}\right) \theta v$. Punishment is greater: when player 2 considers it more likely the cost of compliance is low (r is greater); and when player 2 considers it more likely player 1 is unreasonable (q is lower).

Consider the reasons for these two results. In a type (3) equilibrium, unreasonable people ($I_1 = 0$) never comply and reasonable people ($I_1 = 1$) comply only when the cost of

compliance is low ($C = C_{low}$). Hence, if player 2 observes noncompliance, there might be a good excuse – it might be due to a high cost of compliance ($C = C_{high}$) – but it is also possible that there is no good excuse ($C = C_{low}$ and $I_1 = 0$). When player 2 considers it more likely that the cost of compliance is low (r is greater) or when player 2 considers it more likely player 1 is unreasonable (q is lower), player 2 puts less weight on there being a good excuse. Thus, player 2 gets angrier over noncompliance.

Proposition 2 suggests three observations. First, unreasonable people potentially comply in equilibrium – not just those with a sense of duty. When the cost of compliance is not too high ($C_{high} \leq \theta v$), equilibria of type (1) exist in which noncompliance receives sufficient punishment to induce compliance from unreasonable people with probability 1. This result shows that anger has a distinct role in the model from duty, since it generates compliance from players who would not have complied out of a sense of duty. This property of the model may also be important for the persistence of compliance over time. While we take players' views about duty as exogenous throughout the paper, one might imagine that people's sense of duty to comply with rules would erode if those lacking a sense of duty were able to benefit from breaking the rules.

Second, even when a sense of duty hardly motivates compliance at all (D is very low), it still may be possible to obtain compliance. In particular, if $C_{high} \leq \theta v$, compliance (of both reasonable and unreasonable people) can be achieved for any value of $D > 0$. This result is somewhat surprising: even when duty hardly motivates compliance, it still may be sufficient to generate anger over noncompliance and punishment of noncompliance. On the other hand, in order for player 2 to get angry over noncompliance, it is necessary that player 2 feel there is at least some duty to comply. As we saw in Proposition 1, when player 2 feels there is no duty to comply at all, the unique equilibrium is one in which player 2 excuses noncompliance.

Third, player 2 only gets angry with players whose views regarding duty differ from her own ($I_1 \neq I_2$ types) or whose views player 2 suspects might differ. When $C_{low} = C_{high} = \bar{C}$, first players who hold the same views as the second player regarding duty ($I_1 = I_2$ types)

never provoke anger. First players who hold different views ($I_1 \neq I_2$ types), on the other hand, sometimes do provoke anger: they fail to comply in equilibria of type (1) when $\bar{C} > \theta v$ and this angers player 2. When player 2 does not know the cost of compliance ($C_{high} > C_{low}$), player 2 sometimes becomes angry because of suspected rather than actual differences of opinion: in equilibria of type (3), $I_1 = I_2$ types fail to comply when $C = C_{high}$ and this provokes anger. If player 2 knew that player 1 shared her view regarding duty, she would also know that player 1 had a good excuse ($C = C_{high}$); but she suspects player 2 may disagree about duty and lack a good excuse. The finding that anger arises because of disagreement about duty or the suspicion of disagreement has implications for how we think about conflict situations.

Equilibrium Existence

It follows from Proposition 1 that a perfect Bayesian equilibrium always exists when $I_2 = 0$. When $I_2 = 1$, a perfect Bayesian equilibrium does not always exist. However, the following lemma gives an existence condition.

Lemma 2 *If $I_2 = 1$, a perfect Bayesian equilibrium exists if either of the following conditions is satisfied: (1) F is continuous on $[0, \theta v + D]$ and $F(\theta v) < 1$; or (2) $C = \bar{C}$.*[28]

5 Phenomena Explained by the Model

The model accounts for a large number of phenomena. In particular, it captures two important and ubiquitous enforcement problems. We will also discuss two phenomena explained by the paper's theory of anger and punishment.

Legitimacy

By common definition, a rule is "legitimate" if there is a widespread feeling that there is a duty to comply. Rules lacking legitimacy are difficult – if not impossible – to enforce. The

[28] Condition (1) follows from the Intermediate Value Theorem. For a proof, see the Appendix. Condition (2) follows from Propositon 2.

model explains why legitimacy is so important. When a rule completely lacks legitimacy (which corresponds to the case where $I_1 = I_2 = 0$ in the model), there are three effects. One effect is that people will only comply to avoid punishment. A second effect, which follows from Proposition 1, is that observers of noncompliance will excuse it (since $I_2 = 0$). A third effect, consistent with the model but not directly captured by it, is that observers may fear retaliation if they report rule violations to authorities or punish violators directly. The reason is that punishing noncompliance provokes anger when a rule lacks legitimacy: because reasonable observers (defined by player 1 as $I_2 = 0$ types, since $I_1 = 0$) would be expected to excuse noncompliance. Lack-of-legitimacy thus presents a serious problem for the enforcement of rules.

Legitimacy is crucial in many different contexts. Legal scholars have identified it as a key determinant of the ease or difficulty of enforcing laws (see Tyler (1990) and Fagan and Meares (2008)). The difficulty confronting the police in rooting out gang activity provides an illustrative example. The main problem of the police in the inner city is people's reluctance to report gang activity. Fagan and Meares (2008) point to illegitimacy of the law as the key reason.[29] In his classic study of gangs, Martin Sanchez Jankowski quotes one New York police officer as follows: "When we get the community support, we go with it. It is so frustrating because there are some times when gang members commit a crime in the neighborhood, then we come by, but nobody is willing to help. They say they know nothing."[30] According to Sanchez Jankowski, gang survival depends upon such support from the community. He describes, for instance, the case of the Pink Eagles in New York. Expansion of the gang's drug operations left them less time to patrol the neighborhood, with a resultant increase in robberies and assaults in the community. While the community was willing to excuse the gang's violations of the law, they were angered by the gang's failure to adequately protect the neighborhood. As a result, the community began cooperating with the police. Remarkably, in short order, the Pink Eagles dissolved as an organization.

[29] A number of factors, in their view, contribute to the law's illegitimacy in these communities, such as racially disproportionate incarceration rates.

[30] Sanchez Jankowski (1991), p. 256.

Just as legitimacy plays a major role in the enforcement of the law, it also plays a critical role in the enforcement of rules in firms. We see one example in Gouldner's study of the Oscar Center plant of the General Gypsum Company. Prior to the arrival of a new plant manager, the plant had "few rules... and fewer still that were strictly enforced."[31] The central office charged the new plant manager, Vincent Peele, with making reforms. But, because of the previous lax regime, Peele's new rules were seen as illegitimate and he faced resistance at every turn. Consider Peele's unsuccessful attempt to enforce a rule against absenteeism. According to Gouldner: "Supervisors... did, at first, attempt to enforce this rule. Very shortly thereafter, however, they...declared that this rule just could not be enforced."[32] Supervisors could not ascertain whether those who took absences had good reasons for doing so – since coworkers were unwilling to report when they did not. In addition, when absenteeism was punished, workers became extremely angry (this is the third effect of lack-of-legitimacy, mentioned above). When management decreed that those who took days off without permission would be laid off for an equal number of days, the rate of absenteeism did not fall – it rose. According to Gouldner, "when several workers had been penalized... others would deliberately take off without excuses" as a form of retaliation. As a result, "the number of absentees in any team was greater than usual, and the team would be unable to function."[33]

The need for legitimacy serves as a constraint, one which has been omitted from third-party enforcement models and principal-agent theory more generally. Gouldner describes how the General Gypsum Company came to recognize Peele's lack-of-legitimacy as a constraint. Because rules set by the central office had greater legitimacy than Peele's, the company decided to delegate less authority to him. The central office had not lost faith in Peele: instead, they recognized the value to Peele of being able to cite central office rules in motivating workers. The organizational form that resulted was highly centralized and bureaucratic.

[31] Gouldner (1954), p. 51.
[32] Op. cit., p. 142.
[33] Op. cit., p. 151.

Well-intended policies sometimes fail because of lack of appreciation of the constraints imposed by legitimacy. Ostrom (1990) gives, as an example, the effects of forest nationalizations in Thailand, Niger, Nepal, and India intended to prevent overuse. According to standard third-party enforcement models (such as Becker (1968)), these measures should have been helpful since they set up a system of government policing of forest use where none had previously existed. However, contrary to this prediction, they exacerbated the problem. The respective governments were unable to police effectively because villagers did not feel the government had a legitimate right to nationalize the forests. Furthermore, villages had had their own rules intended to protect their forest parcels from overuse. In ending villages' sense of ownership of their parcels, nationalization delegitimized these village rules – thereby aborting the only effective form of policing that had been taking place.

Noncompliance traps

When rules lack legitimacy, observers of noncompliance excuse it, making enforcement difficult. Observers may also excuse noncompliance simply because the rate of noncompliance is high. This leads to the possibility of a "noncompliance trap": a high rate of noncompliance makes enforcement difficult; the difficulty of enforcing compliance sustains a high rate of noncompliance.

Collier (2000) has argued that pervasive corruption is difficult to fight for just such a reason. In an honest society, a corrupt act "gives rise to a stronger sense of indignation... Indignation is the trigger for disclosure, and so if a corrupt act is detected, it is much more likely to be reported to the authorities." There is at least a scattering of evidence in support of Collier's story. In surveys, for example, people justify their own corrupt behavior by citing its pervasiveness (see Rose-Ackerman (2001)).

The model captures Collier's story.[34] Recall from Proposition 2 and the illustrative example that, when the cost of compliance takes an intermediate value, there are multiple equilibria: one type with a high rate of compliance and anger over noncompliance, and

[34] While this paper gives the only theoretical account of Collier's story, there are several other models of corruption traps. See the introduction for a discussion.

another type with a low rate of compliance and no anger over noncompliance. An honest society might correspond to the former type of equilibrium and a corrupt society might correspond to the latter type. In the corrupt society, because everyone engages in corruption (including reasonable people), corruption fails to provoke anger and is not punished. The lack of punishment sustains the high rate of corruption. In an honest society, reasonable people behave honestly, which means that corrupt acts provoke anger and are punished. This punishment sustains the low rate of corruption.[35]

The model suggests a reason why underdeveloped countries would be especially prone to corruption traps. They may initially lack laws against corruption; and, insofar as laws do exist, they may lack effective organizations to enforce them. In terms of the model, this corresponds to a low θ (low ability to punish noncompliance).[36] According to Proposition 2 and the illustrative example, if θ is sufficiently low, there will be a unique equilibrium, with high corruption. In due course, developing countries may create laws and authorities to enforce them, which increases θ. With a higher θ, there may be multiple equilibria: both a high-corruption equilibrium and a low-corruption equilibrium. While the model is not dynamic, we might expect past corruption and past tolerance of corruption to lead to its persistence even after θ increases.

To escape from such a trap, the sense that corruption is "reasonable" must be erased; this can only be accomplished through a "big push" against corruption. Persson, Rothstein, and Teorell (2012) argue that, historically, dramatic, persistent declines in corruption have generally coincided with big push efforts. Examples include the United States, Sweden, and Denmark, in the nineteenth century, and Hong Kong and Singapore, more recently.

Corruption is just one example of a "noncompliance trap" associated with underdevelop-

[35] The model also accounts for Collier's observation in a second way. We might imagine that the cost of being honest is higher when a society is more corrupt. This corresponds to a higher cost of compliance (C) in the model for corrupt societies. For example, winning a government contract without paying a bribe might be possible for a firm in an honest society but impossible in a corrupt society. According to Proposition 2 and the illustrative example, there might be a unique equilibrium in corrupt societies (with high C) in which people engage in corruption (fail to comply) and corrupt acts fail to provoke anger; in contrast, there might be a unique equilibrium in honest societies (with low C) in which people behave honestly (comply) and corrupt acts provoke anger.

[36] θ is higher when there are laws against corruption and authorities to enforce them because corruption can be punished simply by reporting it.

ment. For example, there are also literatures documenting the persistence – and toleration – of tax avoidance, teacher/health-worker absenteeism, and unpunctuality in underdeveloped countries.[37]

Graduated Sanctions

We turn to a phenomenon explained by the paper's theory of anger and punishment. The model accounts for Ostrom's observation that sanctions generally escalate for repeat offenders. Such escalation of punishment is a common feature of the law: according to Roberts (1997), "for as long as countries have had formal legal systems... recidivists have been seen as deserving harsher punishments than crimes by first offenders."[38]

One explanation for graduated sanctions is suggested by Polinsky and Rubinfeld (1991), who show that they can be optimal in a principal-agent setting. In their model, some agents maximize social welfare, while others have a desire to violate rules when it is not socially optimal. The principal uses graduated sanctions because they permit occasional noncompliance from agents who are appropriately motivated (as is optimal) while still keeping in check those who are not.[39]

[37] Consistent with noncompliance traps, countries with low rates of tax compliance are typically more tolerant of noncompliance (see Lederman (2003)) and find enforcement more difficult (see, for example, Das-Gupta, Lahiri, and Mookherjee (1995)). Similarly, Chaudhury, Hammer, Kremer, Muralidharan, and Rogers (2006) find high tolerance of absenteeism in countries where it is common. In government-run schools in India, for example, only one in three thousand teachers is dismissed for absenteeism annually, despite a 25 percent teacher-absence rate. There is also greater tolerance of unpunctuality when it is common (see Levine, Reis, and West (1980)). Cabral and Pacheco-de-Ameida (2006) cite evidence that the costs of unpunctuality are significant.

[38] In the United States, most states have statutes requiring enhanced punishment for offenders with repeat convictions (see Proband (1994)). California and several other states have enacted "three strikes" laws, which require life sentences for a third serious criminal offense. The United States Supreme Court upheld the constitutionality of a life sentence for a third (and minor) property offense in *Rummel v. Estelle* (see Davis (1992) for a discussion). There is also strong public support for graduated sanctions. In a survey conducted in Missouri, for example, Fichter and Veneziano (1988) found that the percentage of respondents favoring a state prison sentence for an individual who committed a burglary rose from 12 percent on the first offense to 60 percent on the third. Judge William Wilkins, who chaired the United States Sentencing Guidelines Commission, noted that "enhancing a defendant's sentence on the basis of criminal history...is consistent with public perceptions of crime seriousness." (Wilkins (1992), p. 577).

[39] Abreu, Bernheim, and Dixit (2005) find that graduated sanctions can be optimal even if it is never socially optimal for agents to break the rule. Like Polinsky and Rubinfeld (1991), they assume asymmetric information about the agent's cost of compliance. But, they further assume imperfect monitoring of agents, so that an agent suspected of noncompliance may or may not be guilty. Harsh punishment is not meted out for first offenses because of the possibility of type II error.

While Polinsky and Rubinfeld provide one reason for escalating punishment, it has often been suggested that a key reason is that anger is greater when there is a past history of noncompliance. According to Fletcher (1982), "Retributivists hold that, whatever the social utility or disutility of punishment, the recidivist deserves greater punishment."[40] Durham (1987) explains further that "repetitive criminal involvement indicates the existence of 'hidden' attributes possessed by the offender... Personal blameworthiness increases as these hidden features are uncovered."[41]

The model explains Durham's statement. According to Proposition 2, there is more anger over noncompliance when player 2 is more convinced player 1 is unreasonable. A past history of rule violation generally signals that someone is unreasonable. In particular, in a finitely-repeated version of the punishment-compliance game in which player 1's cost of compliance (C) is redrawn each round, a past history of rule violation signals that player 1 is unreasonable.[42,43]

Ostrom cites Glick's (1970) study of the ancient *huerta* irrigation system of Valencia as an example where the cost of compliance was redrawn in this manner. There was a set rotation order for the receipt of water. A farmer was permitted to draw as much water as needed on his turn, but none at other times. According to Ostrom, "from time to time, the cost to a farmer of waiting for his next legal turn to receive water, as contrasted to stealing water available in the canal, would be extraordinarily high."[44] In other words, while most of the time, the cost of compliance (C) was low, occasionally it was extremely high and hence there was a good excuse for taking water. For farmers who stole only in these rare

[40] Fletcher (1982), p.55.

[41] Durham (1987), p. 622.

[42] Durham also suggests that a reasonable person may fail to comply initially from lack of knowledge of the rules. But, upon erring and learning the rules, there is no longer a good excuse for noncompliance. Hence, repeated noncompliance signals that a person is unreasonable. Durham writes: "The offender has been alerted by his previous conviction to the unacceptability of his behavior. The initial conviction, and perhaps any ensuing punishment, acts as an informational conveyance. After being subjected to such an informational barrage, the offender cannot fail to understand the message. Continued criminal behavior therefore reflects a defiance of the law." (p. 621)

[43] One application is to price setting by firms. Rotemberg (2005) has argued that consumers' anger depends upon the frequency of price adjustments.

[44] Ostrom (1990), p. 75.

instances, the "monetary fine...would be quite low."[45]

While the perfect experiment has not yet been run, Kliemann, Young, Scholz, and Saxe (2008) provide suggestive evidence that anger escalates with repeated noncompliance. In their experiment, subjects played two rounds of a trust game with an anonymous opponent.[46] Subjects were then presented with a vignette regarding the opponent. They were told that he had found a neighbor's sweater in his building's washing machine; he moved the sweater to the dryer; and it shrunk four sizes. Subjects assigned greater blame when the opponent had previously failed to reciprocate in the trust game.[47]

Tolerance of self-interestedness in markets

It has been observed that what makes people angry depends upon the context (see Konow (2003) Section 5.2 for a discussion). An example of particular economic significance is the high tolerance of self-interestedness in markets relative to other settings. Adam Smith saw self-interestedness as a principal characteristic: "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest."[48] Sen (1977) argues that economists are too quick, generally, to apply an assumption of self-interestedness in non-market settings. He makes note of the greater tolerance of it in markets – and points out that this difference is a puzzle in need of explanation.

Our model presents a resolution. Depending upon the context, a reasonable person may be defined differently. In a market setting, there is typically a feeling of entitlement to pursue self-interest ($I_1 = I_2 = 0$). According to Proposition 1, when $I_1 = I_2 = 0$, player 1 will pursue self-interest (choose $a = 0$); and this will be tolerated by player 2. In other

[45] Op. cit., p. 75.

[46] Unfortunately, the opponent in this experiment is fictitious.

[47] The experimental setting in Masclet, Noussair, Tucker, and Villeval (2001) would seem to be ideal for identifying whether anger escalates with repeated noncompliance. In their experiment, four subjects repeatedly play a two-stage game. The first stage is a public goods game. In the second stage, subjects have the opportunity anonymously to punish other subjects at a cost. We would hope to see punishments depend upon past levels of contribution as well as the current level. Unfortunately, Masclet et al. (2001) do not report whether this was the case.

[48] Smith (1776), Ch. 2.

settings, there is likely to be less sense of entitlement.[49]

There is considerable tolerance of self-interestedness in markets – but not complete tolerance. The previously mentioned survey of Kahneman, Knetsch, and Thaler (1986) has identified exceptions. Two examples illustrate. 82 percent of respondents considered it unfair for a hardware store to increase the price of snow shovels from \$15 to \$20 the morning after a snowstorm. 77 percent thought it unfair for a small company to decrease workers' wages by 5 percent when the company was making a profit but high unemployment made it easy to replace current employees with new workers at a lower wage.

The model can account for the exceptions to tolerance of self-interestedness identified by Kahneman, Knetsch, and Thaler (1986). All of the cases in which they find such lack of tolerance have a common feature: one or both of the market participants possess market power. For instance, the hardware store is clearly exercising its market power if it raises the price of snow shovels after a snowstorm. It therefore appears that, while market participants normally feel entitled to pursue self-interest ($I_1 = I_2 = 0$), when there is market power they feel there is a duty not to abuse it.

Kahneman, Knetsch, and Thaler sketched their own model to account for their findings. They say that people do not get angry so long as the gains resulting from a transaction are shared (where gains are defined relative to a reference point). Since only gains need to be shared, the model accounts for some of the tolerance of self-interestedness we see. For example, it explains why the failure of the rich to give away most of their wealth does not necessarily provoke anger. But their model fails to account for competitive market situations (as in the stock market), where there is tolerance of the self-interested pursuit of

[49] Note that the model gives a second reason why self-interestedness may be tolerated in markets. According to Proposition 2 (and the illustrative example of Section 2), when the cost of compliance is high, player 1 will act according to self-interest and player 2 will not get angry. The cost of compliance is often high in markets. For example, in perfectly competitive markets, firms that set prices below the market price make losses and are driven out of business. Hence, even if market participants feel that there is a duty to keep prices low, they may still excuse a firm for charging the market price, since the cost of doing otherwise is so high. Schotter, Weiss, and Zapater (1996) have provided experimental evidence in support of this idea. In their experiment, subjects played an ultimatum game. Those subjects assigned to the first-player role moved on to a second stage, where there was an opportunity to earn more, if their earnings in the first stage were in the top half of the distribution. Schotter, Weiss, and Zapater found that such competition to move on to the second stage made second players more excusing of low offers in the first round.

gain. In contrast, our model, which is quite different from theirs, explains such situations comfortably.

6 Conclusion

This paper elaborates a theory of rule enforcement where people are angered by violations of rules. Corresponding to the standard definition of negligence in the law, noncompliance provokes anger in the model when a "reasonable person" would have complied under similar circumstances.

The model illuminates the circumstances in which rules will be enforceable. It is unique in identifying the "legitimacy" of rules as a determinant of enforceability. While there is a large literature on legitimacy outside of economics, within economics, it has been all-but-unexplored. The model suggests that rules are also difficult to enforce when the rate of noncompliance is high. Collier (2000) has argued that pervasive corruption is hard to fight for exactly this reason. This leads to the possibility of multiple equilibria in the model: one equilibrium in which noncompliance is tolerated and there is a low rate of compliance, and another in which noncompliance provokes anger and there is a high rate of compliance.

The model also yields a theory of anger. This theory explains why anger and punishment normally escalate with repeat offense. The model also accounts for contextual differences in what provokes anger, such as the high degree of tolerance of self-interestedness in markets relative to other settings.

The paper suggests many directions for future research – especially with regard to firms and other organizations. It suggests the possibility of welfare-destroying conflict caused by differing views of duty. How do organizations mitigate internal conflict? Within organizations, there are not only rules as to how people should behave but also rules concerning who should punish misbehavior and to what extent. Can such "institutionalization" of punishment increase its provision? Furthermore, rule violators might be angered if they are punished and retaliate. When does the threat of such retaliation prevent observers

from reporting noncompliance to authorities?[50]

The paper especially raises two important issues. The need to legitimate rules serves as a constraint on the kinds of rules that can be enforced. This constraint is absent from existing third-party enforcement models and principal-agent theory more generally. How do such constraints affect the way in which organizations are structured? (One example – surely just the tip of an iceberg – is the bureaucracy that arose in the General Gypsum Company because the plant manager lacked legitimacy.) Second, our discussion of markets implies that fairness attitudes differ from those within firms; what are the consequences for the theory of the firm? The high tolerance of self-interestedness in markets suggests that a special property of markets is that outcomes are obtained with a minimum of contention. For example, Adam Smith's butcher, brewer, and baker not only provide dinner; they do so without complaint. We might expect a different outcome if, instead, the butcher, brewer, and baker were employees within the same firm. What are the implications for how activity should be distributed between firms and markets?

[50]The model also gives a framework for thinking about social change. Recent history is replete with examples of dramatic shifts in what is tolerated: discrimination (based on race, gender, and sexual orientation); divorce; children out-of-wedlock; skin exposure; language; child-rearing; the list goes on. The model suggests such social changes reflect an altered conception of what is "reasonable." A change in what is considered reasonable might be the result of new ideas (a change in I_1 and I_2); it might also reflect a shift from one equilibrium to another. Social change does not appear to be solely due to changes in ideas. In the case of the Civil Rights Movement, for example, tolerance of discrimination appears to have declined quickly while deep-seated attitudes about race changed at a slower pace (see Sniderman and Tetlock (1986)).

References

Abreu, Dilip, B. Douglas Bernheim, and Avinash Dixit. 2005. "Self-Enforcing Cooperation with Graduated Punishments." Working Paper.

Acemoglu, Daron. 1995. "Reward structures and the allocation of talent." *European Economic Review* 39: 17–33.

Akerlof, George and Rachel Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19: 9-32.

Andvig, Jens and Karl Moene. 1990. "How Corruption May Corrupt." *Journal of Economic Behavior and Organization* 13: 63-76.

Becker, Gary. 1968. "Crime and Punishment: An Economic Approach." *The Journal of Political Economy* 76: 169–217.

Benabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70: 489–520.

Benabou, Roland and Jean Tirole. 2011. "Identity, Morals and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics* 126: 805-855.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102(5): 841-877.

Besley, Timothy and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents" *American Economic Review* 95(3): 616-36.

Blau, Peter M. 1964. *Exchange and Power in Social Life*. New York, NY: Wiley.

Cabral, Luis, and Gonçalo Pacheco-de-Almeida. 2006. "Punctuality: A Research Agenda." Working Paper.

Cadot, Olivier. 1987. "Corruption as a Gamble." *Journal of Public Economics* 33: 223-244.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20(1): 91-116.

Chen, Daniel. 2013. "The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I." NBER Working Paper.

Coleman, James S. 1990. *Foundations of Social Theory.* Cambridge, MA: Harvard University Press.

Collier, Paul. 2002. "How to Reduce Corruption." *African Development Review* 12(2): 191-205.

Das-Gupta, Arindam, Radhika Lahiri, and Dilip Mookherjee. 1995. "Income Tax Compliance in India: An Empirical Analysis." *World Development* 23(12): 2051-2064.

Davis, M. 1992. "Just Deserts for Recidivists." In *To Make the Punishment Fit the Crime: Essays in the Theory of Criminal Justice*. Boulder, CO: Westview.

Durham, Alexis. 1987. "Justice in Sentencing: The Role of Prior Record of Criminal Involvement." *The Journal of Criminal Law and Criminology* 78(3): 614-43.

Elster, Jon. 1989. "Social norms and economic theory." *Journal of Economic Perspectives* 4: 99-117.

Fagan, Jeffrey and Tracey L. Meares. 2008. "Punishment, Deterrence, and Social Control: The Paradox of Punishment in Minority Communities." *Ohio State Journal of Criminal Law* 6: 173-229.

Fehr, Ernst, Urs Fischbacher, and Simon Gächter. 2002. "Strong Reciprocity, Human Cooperation, the Enforcement of Social Norms." *Human Nature* 13(1): 1-25.

Fletcher, George. 1982. "The Recidivist Premium." *Criminal Justice Ethics* 1: 54-59.

Glick, Thomas F. 1970. *Irrigation and Society in Medieval Valencia*. Cambridge, MA: Harvard University Press.

Gouldner, Alvin W. 1954. *Patterns of Industrial Bureaucracy*. New York, NY: The Free Press.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* 76: 728-741.

Kandori, Michihiro. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59: 63-80.

Konow, James. 2003. "Which Is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature* 41(4): 1188-1239.

Kliemann, Dorit, Liane Young, Jonathan Scholz, and Rebecca Saxe. 2008. "The Influence of Prior Record on Moral Judgment." *Neuropsychologia* 46: 2949-2957.

Lederman, Leandra. 2003. "The Interplay Between Norms and Enforcement in Tax Compliance." *Ohio State Law Journal* 64(6): 1453-1514.

Levine, David. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1: 593-622.

Levine, Robert, Laurie West, and Harry Reis. 1980. "Perceptions of Time and Punctuality in the United States and Brazil." *Journal of Personality and Social Psychology* 38(4): 541-550.

Lindbeck, Assar, Sten Nyberg, and Jörgen Weibull. 1999. "Social norms and economic incentives in the welfare state." *Quarterly Journal of Economics* 114: 1-35.

Lui, Francis. 1985. "An Equilibrium Queuing Model of Bribery." *Journal of Political Economy* 93(4): 760-81.

Masclet, David, Charles Noussair, Steven Tucker, Marie-Claire Villeval. 2001. "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review* 93(1): 366-380.

Miller, Alan D. and Ronen Perry. 2012. "The Reasonable Person." *New York University Law Review* 87(2): 323-92.

Murphy, Kevin M., Andrei Shleifer and Robert Vishny. 1991. "The allocation of talent: implications for growth." *Quarterly Journal of Economics* 106: 503–30.

Murphy, Kevin M., Andrei Shleifer and Robert Vishny. 1993. "Why is rent seeking so costly to growth?" *American Economic Review Paper and Proceedings* 83: 409–14.

Nikiforakis, Nikos, Charles N. Noussair, and Tom Wilkening. 2012. "Normative conflicts and feuds: The limits of self-enforcement." *Journal of Public Economics* 96(9-10): 797-807.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Persson, Anna, Bo Rothstein, and Jan Teorell. 2012. "Why Anticorruption Reforms Fail – Systemic Corruption as a Collective Action Problem." *Governance* (forthcoming).

Polinsky, A. Mitchell and Daniel Rubinfeld. 1991. "A Model of Optimal Fines for Repeat Offenders." *Journal of Public Economics* 46: 291-306.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97(1): 180-96.

Prendergast, Canice. 2008. "Intrinsic Motivation and Incentives." *American Economic Review Papers and Proceedings* 98(2): 201-5.

Proband, Stan C. 1994. "Habitual Offender Laws Ubiquitous." *Overcrowded Times* 5(1): 7.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281-1302.

Roberts, Julian. 1997. "The Role of Criminal Record in the Sentencing Process." *Crime and Justice* 22: 303-362.

Rose-Ackerman, Susan. 2001. "Trust and Honesty in Post-Socialist Societies." *Kyklos* 54: 415–44.

Rotemberg, Julio. 2005. "Customer Anger at Price Increases, Changes in the Frequency of Price Adjustment and Monetary Policy." *Journal of Monetary Economics* 52(4): 829–52.

Sah, Raaj. 1988. "Persistence and Pervasiveness of Corruption: New Perspectives." Yale Economic Growth Center Discussion Paper 560.

Sanchez Jankowski, Martin. 1991. *Islands in the Street*. London, UK: University of California Press.

Schotter, Andrew, Avi Weiss, and Inigo Zapater. 1996. "Fairness and Survival in Ultimatum and Dictatorship Games." *Journal of Economic Behavior and Organization* 31: 37-56.

Sen, Amartya. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6(4): 317-44.

Smith, Adam. 1776. *The Wealth of Nations*. Edited by Edwin Cannan, 1904. Reprint edition, 1937. New York, NY: Modern Library.

Sugden, Robert. 1986. *The Economics of Rights, Cooperation, and Welfare*. Oxford, UK: Basil Blackwell.

Sniderman, Paul M. and Philip E. Tetlock. 1986. "Symbolic Racism: Problems of Motive Attribution in Political Analysis" *Journal of Social Issues* 42(2): 129-150.

Tyler, Tom. 1990. *Why People Obey the Law*. New Haven, CT: Yale University Press.

Unikel, Robert. 1992. "'Reasonable' Doubts: A Critique of the Reasonable Woman Standard in American Jurisprudence." *Northwestern University Law Review* 87(1): 326-375.

United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947).

Wilkins, William W., Jr. 1992. "The Federal Sentencing Guidelines: Striking an Appropriate Balance." *University of California Davis Law Review* 25: 571-86.

Young, H. Peyton. 1993. "The Evolution of Conventions." *Econometrica* 61(1): 57-84.

Young, H. Peyton. 1996. "The Economics of Convention." *Journal of Economic Perspectives* 10(2): 105-122.

Young, H. Peyton. 2008. "Social Norms." In *The New Palgrave Dictionary of Economics, Second Edition*, eds. Steven N. Durlauf and Lawrence E. Blume. London: Macmillan.

7 Appendix (For Online Publication)

7.1 Reasonable-Person (RP) Stability

Our main focus in the paper is on perfect Bayesian equilibria. In Section 2's illustrative example, however, we used a different equilibrium concept because its conditions were easier to state. This equilibrium concept corresponds to a refinement of perfect Bayesian equilibrium. To show the correspondence between the illustrative example and the formal model, we will now define this refinement. This refinement eliminates only a few PBE. It generates the same compliance behavior as PBE, but PBE admits a larger set of possible punishments of noncompliance.

We will say that a perfect Bayesian equilibrium is reasonable-person (RP) stable if player 2's beliefs about reasonable people are "stable" in the following sense. If player 2 is certain ex ante that reasonable players always comply, player 2 remains certain ex post. RP stability is defined more formally below. Note that RP stability is not a standard refinement concept. It requires that player 2 hold stable beliefs about how reasonable people behave, where reasonableness is a concept particular to the game considered in this paper.

Definition 2 *A perfect Bayesian equilibrium is reasonable-person (RP) stable if:*

(D4) $\mu(\{C : \tilde{a}^{rp^*}(C) = 1\}|0) = 1$ *whenever* $\mu(\{C : \tilde{a}^{rp^*}(C) = 1\}) = 1$, *where* $\tilde{a}^{rp^*}(C) = \tilde{a}^*(C, I_2)$.

If player 2 expects noncompliance to arise with positive probability, so that Bayes' rule is applicable ($F(\tilde{p}^*) < 1$), condition D4 is implied by condition D3. If, on the other hand, player 2 never expects to see noncompliance, so that Bayes' rule is not applicable, condition D4 places an additional restriction on player 2's posterior beliefs ($\mu(\cdot|0)$). It requires that player 2 believe, upon observing noncompliance, that reasonable types nonetheless always comply. Clearly, player 2 will feel mistreated in amount v in such a circumstance, since she observed noncompliance but believes a reasonable person would have complied ($E_{\mu(\cdot|0)}[M(\tilde{a}^*, C, I_1)] = v$). It follows from condition (*) that noncompliance will receive punishment $\tilde{p}^* = \theta v$. This is stated formally in the following lemma.

Lemma 3 *A perfect Bayesian equilibrium is RP-stable if and only if:*

(P4) *If* $F(\tilde{p}^*) = 1$*:* $\tilde{p}^* = \theta v$.

The following are restatements of Propositions 1 and 2, which characterize the RP-stable equilibria in addition to the perfect Bayesian equilibria.

Proposition 1 (restatement) *If player 2 feels there is no duty to comply* ($I_2 = 0$)*, a unique perfect Bayesian equilibrium exists and it is RP-stable. In this equilibrium, player 2 does not punish noncompliance:*

$\tilde{p}^* = 0$

$$\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$$

Proposition 2 (restatement) *Suppose player 2 feels there is a duty to comply ($I_2 = 1$), and player 2 believes $C = C_{low} > 0$ with probability r and $C = C_{high} \geq C_{low}$ with probability $1 - r$ $(0 < r < 1)$. There are three types of perfect Bayesian equilibria that can arise.*[51]

Type 1: *the reasonable type ($I_1 = 1$) always complies.*

If $C_{high} \leq D + \theta v$, PBE exist with:

(i) $\tilde{p}^* \in [\min(\theta v, C_{high}), \theta v]$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p}^* + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

The equilibrium with $\tilde{p}^* = \theta v$ is RP-stable.

Type 2: *the reasonable type never complies.*

If $C_{low} > D$, a PBE exists with:

(i) $\tilde{p}^* = 0$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

This equilibrium is RP-stable.

Type 3: *the reasonable type sometimes complies (when $C = C_{low}$).*

If $\left(\frac{r-rq}{1-rq}\right)\theta v < C_{low} \leq D + \left(\frac{r-rq}{1-rq}\right)\theta v < C_{high}$, a PBE exists with:

(i) $\tilde{p}^* = \left(\frac{r-rq}{1-rq}\right)\theta v$

(ii) $\tilde{a}^*(C, I_1) = \begin{cases} 1, \text{ if } C \leq \left(\frac{r-rq}{1-rq}\right)\theta v + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$

This equilibrium is RP-stable.

Observe that the RP-stable equilibria in Propositions 1 and 2 are nearly identical to the full set of perfect Bayesian equilibria. The only difference is that, in Proposition 2, there is a unique RP-stable equilibrium of Type 2 while there are multiple perfect Bayesian equilibria. The RP-stable equilibria in Proposition 2 are identical to the equilibria in Figure 1. As mentioned previously, the full set of perfect Bayesian equilibria produce a nearly identical picture. The only difference relative to Figure 1 is that a wider range of punishments are admissible in Region 1A: $\tilde{p}^* \in [\min(\theta v, C_{high}), \theta v]$.

From the restatement of Proposition 1, it follows that an RP-stable equilibrium always exists when $I_2 = 0$. When $I_2 = 1$, an RP-stable equilibrium does not always exist. However, Lemma 4 gives existence conditions (they are, in fact, the same existence conditions as those given for PBE in Lemma 2).

Lemma 4 *If $I_2 = 1$, an RP-stable perfect Bayesian equilibrium exists if either of the following conditions is satisfied: (1) F is continuous on $[0, \theta v + D]$ and $F(\theta v) < 1$; or (2) $C = \bar{C}$.*[52]

[51] If $C_{low} \leq \min\left(D, \left(\frac{r-rq}{1-rq}\right)\theta v\right)$ and $C_{high} > D + \theta v$, an equilibrium does not exist. This is consistent with Lemma 3 (see below), since Lemma 3 does not ensure existence in this particular case.

[52] Condition (1) follows from the Intermediate Value Theorem. For a proof, see the Appendix. Condition (2) follows from Propositon 2.

7.2 Proofs

Proof of Lemma 1. Condition P1 is established in Section 3. Section 3 also discusses the case where $I_2 = 0$. So, we will restrict attention to establishing conditions P2 and P3 for the case where $I_2 = 1$.

First, let us consider when Bayes' rule is applicable. It will be applicable whenever $a = 0$ arises with positive probability. Since, according to our restatement of condition D1, the $I_1 = 0$ type is less compliant than the $I_1 = 1$ type, $a = 0$ with positive probability if and only if the $I_1 = 0$ type chooses $a = 0$ with positive probability. From our restatement of condition D1, we know that the $I_1 = 0$ type chooses $a = 0$ if $\tilde{p}^* < C$. Thus Bayes' rule applies so long as $F(\tilde{p}^*) < 1$.

Now, consider condition (*). It requires that:

$$\begin{aligned} \tilde{p}^* &= \theta \cdot E_{\mu(\cdot|0)}[M(\tilde{a}^*, C, I_1)] \\ &= \theta v \cdot E_{\mu(\cdot|0)}[\max(\tilde{a}(C, I_2) - \tilde{a}(C, I_1), 0)] \end{aligned}$$

If we apply our restatement of condition D1, it follows that:

$$\tilde{p}^* = \theta v \cdot \mu(\{(C, I_1) : \tilde{p}^* + I_1 \cdot D < C \leq \tilde{p}^* + I_2 \cdot D\}|0)$$

When $F(\tilde{p}^*) = 1$, Bayes' rule is not applicable so there are no restrictions on player 2's posterior beliefs ($\mu(\cdot|0)$). If $I_2 = 1$, we can choose $\mu(\cdot|0)$ so that $\mu(\{(C, I_1) : \tilde{p}^* + I_1 \cdot D < C \leq \tilde{p}^* + I_2 \cdot D\}|0)$ takes any value in the interval $[0, 1]$. Thus, when $F(\tilde{p}^*) = 1$, $\tilde{p}^*$ can take any value in the interval $[0, \theta v]$.

When Bayes' rule is applicable and $I_2 = 1$:

$$\begin{aligned} \tilde{p}^* &= \theta v \cdot \mu(\{(C, I_1) : \tilde{p}^* + I_1 \cdot D < C \leq \tilde{p}^* + D\}|0) \\ &= \theta v \cdot \mu(\{(C, I_1) : \tilde{p}^* < C \leq \tilde{p}^* + D \text{ and } I_1 = 0\}|0) \\ &= \theta v \cdot \frac{\mu(\{(C, I_1) : \tilde{p}^* < C \leq \tilde{p}^* + D \text{ and } I_1 = 0\})}{\mu(\{(C, I_1) : \tilde{a}^*(C, I_1) = 0\})} \\ &= \theta v \frac{\mu(\{(C, I_1) : \tilde{p}^* < C \leq \tilde{p}^* + D \text{ and } I_1 = 0\})}{\mu(\{(C, I_1) : \tilde{p}^* + I_1 \cdot D < C\})} \\ &= \theta v \cdot \frac{[F(\tilde{p}^* + D) - F(\tilde{p}^*)](1 - q)}{[1 - F(\tilde{p}^*)] - q[F(\tilde{p}^* + D) - F(\tilde{p}^*)]} \end{aligned}$$

This completes the proof. ■

Proof of Lemma 2. Lemma 2 follows immediately from Lemma 4, so it is sufficient to prove Lemma 4. ■

Proof of Lemma 3. A proof of Lemma 3 is omitted, since it is given in the text. ■

Proof of Lemma 4. Existence is established in case (2) by Proposition 2. Therefore, we only need to establish existence in case (1). Let us consider the set of strategy pairs $(\tilde{a}(C, I_1), \tilde{p})$ with the properties that (1) $\tilde{a}(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p} + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$ and (2) $\tilde{p} \in$

$[0, \theta v]$. We can index the members of this set by the value of $\tilde{p}$. We will prove existence by showing that some strategy pair in this set satisfies the conditions P1 through P4. First, observe that condition P1 is trivially met for all members of the set. Since $F(\theta v) < 1$, $F(\tilde{p}) < 1$ for all $\tilde{p}$ in the set. Hence, condition P2 is applicable while conditions P3 and P4 are not. Observe that P2 holds if and only if $\tilde{p} = \theta v \cdot \frac{[F(\tilde{p}+D)-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[F(\tilde{p}+D)-F(\tilde{p})]}$. When $\tilde{p} = 0$, the left-hand-side of this equation is zero and the right-hand-side is greater than or equal to zero: $LHS \leq RHS$. When $\tilde{p} = \theta v$, $RHS \geq LHS$. Since F is continuous on $[0, \theta v + D]$, the Intermediate Value Theorem implies that there exists $\tilde{p} \in [0, \theta v]$ satisfying condition P2. This proves existence. ■

Proof of Proposition 1. Suppose $I_2 = 1$. From Lemma 1, we know that if a PBE exists, it must have $\tilde{p} = 0$ and $\tilde{a}(C, I_1) = \begin{cases} 1, \text{ if } C \leq D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$. This shows uniqueness. Now let us show existence. The strategy pair clearly meets condition P1 of Lemma 1. $F(\tilde{p}) = F(0) < 1$ since C is assumed to always be greater than zero. Hence, condition P2 is applicable while conditions P3 and P4 are not applicable. Condition P2 is clearly satisfied. Lemma 3 also implies that the equilibrium is RP-stable, since condition P4 does not apply. This completes the proof. ■

Proof of Proposition 2. Let us consider the set of strategy pairs $(\tilde{a}(C, I_1), \tilde{p})$ with the properties that (1) $\tilde{a}(C, I_1) = \begin{cases} 1, \text{ if } C \leq \tilde{p} + D \cdot I_1 \\ 0, \text{ otherwise} \end{cases}$ and (2) $\tilde{p} \in [0, \theta v]$. We can index the members of this set by the value of $\tilde{p}$. Lemma 1 implies that a PBE must be drawn from this set.

In a perfect Bayesian equilibrium, one of the following must be true: (i) an $I_1 = 1$ type always complies, (ii) an $I_1 = 1$ type never complies, or (iii) an $I_1 = 1$ type complies only when $C = C_{low}$.

Consider case (i). If $\tilde{p}$ is a PBE of type (i), we must have $C_{high} \leq \tilde{p} + D$. First, suppose $C_{high} \leq \tilde{p}$ (we will refer to this as case (i)-A). This implies that $F(\tilde{p}) = 1$, so condition P2 is not applicable but conditions P1, P3, and P4 are applicable. Condition P1 is clearly satisfied for any strategy pair in the set. Condition P3 requires $\tilde{p} \in [0, \theta v]$ while condition P4 requires $\tilde{p} = \theta v$. This proves that, if a PBE of type (i)-A exists, it must be a strategy pair with $\tilde{p} \in [0, \theta v]$ (and furthermore, if it exists, it will be RP-stable if $\tilde{p} = \theta v$). Existence requires that $C_{high} \leq \tilde{p}$, or $\tilde{p} \in [\min(\theta v, C_{high}), \theta v]$. Existence also requires $C_{high} \leq \tilde{p} + D$, or $C_{high} \leq \theta v + D$.

Now consider case (i)-B, in which $\tilde{p} < C_{high} \leq \tilde{p} + D$. This implies $F(\tilde{p}) < 1$, so conditions P1 and P2 are applicable while P3 and P4 are not applicable. Condition P1 is clearly satisfied for any strategy pair in the set. Condition P2 requires that $\tilde{p} = \theta v \cdot \frac{[F(D+\tilde{p})-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[F(D+\tilde{p})-F(\tilde{p})]}$. Since $F(D + \tilde{p}) = 1$, it follows that $\tilde{p} = \theta v \cdot \frac{[1-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[1-F(\tilde{p})]} = \theta v$. This proves that, if a perfect Bayesian equilibrium of type (i)-B exists, it must be the strategy pair with $\tilde{p} = \theta v$ (and furthermore, if it exists, it will be RP-stable). The strategy pair with $\tilde{p} = \theta v$ will indeed be an RP-stable PBE if $\theta v < C_{high} \leq \theta v + D$.

Combining cases (i)-A and (i)-B, we conclude $\tilde{p} \in [\min(\theta v, C_{high}), \theta v]$ will be a perfect Bayesian equilibrium of type (i) if $C_{high} \leq \theta v + D$; it will be RP-stable if $\tilde{p} = \theta v$.

Now, consider case (ii). If $\tilde{p}$ is a PBE of type (ii), we must have $C_{low} > \tilde{p} + D$. This implies that $F(D + \tilde{p}) = F(\tilde{p}) = 0$. Hence, only conditions P1 and P2 of Lemma 1 are

applicable. Condition P1 is clearly satisfied for any strategy pair in the set. Condition P2 is met if: $\tilde{p} = \theta v \cdot \frac{[F(D+\tilde{p})-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[F(D+\tilde{p})-F(\tilde{p})]} = 0$. This proves that, if a PBE of type (ii) exists, it must be the strategy pair with $\tilde{p} = 0$ (and furthermore, if it exists, it will be RP-stable). The strategy pair with $\tilde{p} = 0$ will indeed be a PBE of type (ii) if $C_{low} > \tilde{p} + D$, or $C_{low} > D$.

Finally, consider case (iii). If $\tilde{p}$ is a PBE of type (iii), we must have $C_{low} \leq \tilde{p} + D < C_{high}$. This implies that $F(D + \tilde{p}) = r < 1$. Since $F(\tilde{p}) < F(D + \tilde{p}) < 1$, conditions P1 and P2 are applicable while conditions P3 and P4 are not. P1 is clearly satisfied for any strategy pair in the set. Condition P2 requires that $\tilde{p} = \theta v \cdot \frac{[F(D+\tilde{p})-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[F(D+\tilde{p})-F(\tilde{p})]}$. Since $F(D + \tilde{p}) = r$, $\tilde{p} = \theta v \cdot \frac{[r-F(\tilde{p})](1-q)}{[1-F(\tilde{p})]-q[r-F(\tilde{p})]}$. If $C_{low} \leq \tilde{p}$, in which case $F(\tilde{p}) = r$, $\tilde{p} = \theta v \cdot \frac{[r-r](1-q)}{[1-r]-q[r-r]} = 0$. But, if $\tilde{p} = 0$, it cannot be the case that $C_{low} \leq \tilde{p} = 0$. Hence, we must have $C_{low} > \tilde{p}$, in which case $F(\tilde{p}) = 0$. Since $F(\tilde{p}) = 0$, $\tilde{p} = \theta v \cdot \frac{[r-0](1-q)}{[1-0]-q[r-0]} = \theta v \cdot \frac{r-rq}{1-rq}$. This proves that, if a PBE of type (iii) exists, it must be the strategy pair with $\tilde{p} = \theta v \cdot \frac{r-rq}{1-rq}$ (and furthermore, if it exists, it will be RP-stable). In order for this strategy pair to indeed be a PBE of type (iii), we must have $C_{low} \leq \tilde{p} + D < C_{high}$, or $C_{low} \leq \theta v \cdot \frac{r-rq}{1-rq} + D < C_{high}$. We must also have $C_{low} > \tilde{p}$, or $C_{low} > \theta v \cdot \frac{r-rq}{1-rq}$. Hence, if a type (iii) PBE exists, it will be the strategy pair with $\tilde{p} = \theta v \cdot \frac{r-rq}{1-rq}$ and such an equilibrium exists when $\theta v \cdot \frac{r-rq}{1-rq} < C_{low} \leq \theta v \cdot \frac{r-rq}{1-rq} + D < C_{high}$. This completes the proof. ■