Is Full Compliance Possible? Conditions for Shirking with Imperfect Monitoring and Continuous Action Spaces

Jenna Bednar*
Department of Political Science
University of Michigan
jbednar@umich.edu

January 2006

Abstract

Games of public good provision, collective action, and collusion share concern for the free rider that shirks on its obligations. According to the folk theorem, the free rider problem can be resolved through punishment mechanisms. Versions of the folk theorem have been applied when monitoring is imperfect. Empirical evidence contradicts this theory: while often subjects cooperate significantly, rarely is all shirking eliminated. To reconcile theory with empirical evidence, I provide a theoretical basis for the inability of participants in collective action problems to attain full compliance. I construct a general class of compliance models with imperfect monitoring through a common signal. I derive sufficient conditions—on both the utility of agents and the monitoring capabilities—under which slippage from full compliance is unavoidable, showing the limits of the folk theorem logic. The results cover most cases of concern to political scientists and political economists including public goods provision, contract and treaty compliance, collective action, and even Cournot competition. The paper concludes with a discussion about institutional design.

The free rider problem—fundamental to the study of politics—exists in games of compliance, collective action, public good provision, and collusion. These games often share similar characteristics: each participant gains from the sacrifices made by others but prefers not to make the sacrifices himself, and yet prefers full sacrifice (by everyone, including himself) to

*I thank the many who have suggested improvements to this paper, including Bob Axelrod, James Bowers, Sven Feldmann, Cathy Hafer, Ken Kollman, Skip Lupia, Burt Monroe, Jim Morrow, and Scott Page.

complete non-compliance. We are concerned about the shirker when we study intergovernmental opportunism in international treaties (e.g. Downs & Rocke 1995, Downs et al 1996, Fearon 1998, Simmons 2000, Morrow 2001) and within federal systems (de Figueiredo and Weingast 2005, Bednar 1998, Dougherty 2001), and individual opportunism in commons problems (Hardin 1968, Ostrom 1990), contract compliance (Greif, Milgrom & Weingast 1994), tax compliance (Scholz & Lubell 1998), mass movements, including labor strikes (Olson 1965), establishment of the rule of law (Weingast 1997), management of ethnic conflict (Fearon & Laitin (1996), misrepresentation of ethnic kinship (Axelrod & Hammond 2003), and collusion in cartels (Green & Porter 1984).

In this paper, I consider the free rider problem in environments where actions are not fully verifiable. As a specific example, consider international trade agreements. The visibility of trade barriers varies; certain restrictions, such as tariffs or quotas on foreign imports, are easy to spot, while other domestic policies, such as industry support, affect trade competitiveness but are more difficult to identify or categorize as a barrier to trade. While a major improvement of the WTO over the GATT is to take on the deterrence of non-tariff barriers, its dispute resolution panel is active. Despite incentives to deter trade barriers, evidence of minor non-compliance is widespread. The inability to sustain full compliance is not unique to the WTO; it is a problem in every one of the above contexts. Despite full compliance's tantalizing draw,[1] its rarity is unexplained, as a general phenomenon, by the theoretical compliance literature.

Compliance when actions cannot be perfectly observed is an essential problem of political science. To explain it, theorists initially focused on the possibility of compliance. This literature, largely spawned by Axelrod (1984) and now known as Cooperation Theory, shows that when these games are repeated, the free rider problem can be overcome. While the models are constructed to pursue the problem of *existence* of cooperation, not the *extent* of

[1]Within political realms its focality is established for three reasons: it is often easier to discern deviations from full compliance in political settings (in an economic environment where the signal is price, the profit-maximizing price is no more or less transparent than any other price); legally, it makes it easier to define an agreement; and the relationship may be stronger if the participants believe that all others are complying fully and not deviating by small margins.

it, invocations of the folk theorem imply the existence of a full compliance equilibrium.[2]

Empirical evidence contradicts this inference. While advances in behavioral economics and political science suggest that people are more cooperative than we would often anticipate in single-shot or finitely-repeated games, subjects also fail to cooperate fully in repeated settings despite motivating incentives.[3] Summarizing hundreds of laboratory results, and armed with decades of field experience, Ostrom writes (1999:508): "[T]he temptation to cheat always exists. No amount of monitoring and sanctioning reduces the temptation to zero."

To reconcile theory and empirical evidence, this paper inverts the original puzzle of Cooperation Theory to ask: Under what conditions is full compliance possible? I construct a general model of homogenous agent compliance with continuous action spaces and imperfect monitoring through a common signal that allows for arbitrary compliance levels. Using this model, I provide sufficient conditions—on both the utility of agents and the monitoring capabilities—under which slippage from full compliance is unavoidable. My results cover

[2]It must be underscored that the folk theorem only implies that full compliance is possible; for many strategic contexts, it is only one of many potential equilibria. Nevertheless, of all equilibria supported, it is full compliance that catches the most attention. A product of theory's prediction of full compliance is that institutional design presumes that full compliance is a viable, and complete, objective. In a special issue of *International Organization* dedicated to the rational design of institutions, the editors write: "The cooperation literature is premised on the 'Folk Theorem,' which shows that cooperation is possible in repeated games. ...In brief, the possibility of cooperation is present in most modern international issues" (Koremenos, Lipson, and Snidal 2001:764-65). To explain the realization of partial compliance, rather than full, several papers within the volume introduce modifications to the standard compliance game, including agent heterogeneity and uncertainty about payoffs.

[3]Public goods experiments identify two types of disagreements between game theory and laboratory results: subjects contribute both more and less than predicted (Ledyard 1995, Palfrey and Prisbrey 1997). While I know of no experiment that combines continuous choice and imperfect information, some results are suggestive. When researchers have intentionally designed the experiment to encourage cooperative behavior, full compliance is never achieved. Feinberg and Snyder (2002) let subjects choose three prices; with price uncertainty, subjects fail to coordinate on the utility-maximizing collusive price. Continuous choice experiments are equally without success in obtaining full cooperation, despite instituting communication devices (Ostrom et al 1992), unanimity rules (Banks, Plott, and Porter 1988), and even making contribution the dominant strategy (Saijo and Nakamura 1995, Palfrey and Prisbey 1997).

In field research, examples of the inability to obtain full compliance abound. In international law, it is generally assumed that in treaties, governments will shirk slightly. For example, while Chayes and Chayes (1993) champion the high degree of treaty compliance, they acknowledge that compliance is partial. Full compliance is an unrealistic standard (1993:176). See also Brown Weiss and Jacobson 1998 and Simmons 2000.

most cases of concern to political scientists and political economists including public goods provision, contract and treaty compliance, collective action, and even Cournot competition.

My model builds upon binary choice models in which agents choose either to comply or not comply such as Abreu, Pierce and Stachetti (1986, 1990). Though binary choice models enable us to show that compliance can be sustained, they do not allow for slippage from full compliance. And this slippage, for reasons that I will discuss, has political and legal implications. The work that is perhaps closest to this model is the seminal paper by Green and Porter (1984). They solve for optimal oligopolistic production in which signals are prices. Since they focus on deriving a profit maximizing equilibrium, they do not rule out a full compliance equilibrium that generates lower utility. This paper has a different emphasis: it sets out to understand the degree of compliance. I show that in general *any* equilibrium involves slippage from full compliance. This result is proven for classes of utility functions, not just for a specific functional form, and allows for monitoring mechanisms other than prices, which rarely exist in political contexts.

Full compliance may be possible when an agent is indifferent between a limited set of alternatives, or when the rules define something it wants to do, or can't help but do, already. Full compliance is also possible when actions are perfectly visible to others or when the punishment is disproportionate to the crime: death for running a stop sign. But this paper will show that when we are asked to do something that is costly for us to do, and we know that others can't see our action perfectly, we shirk a little bit.[4] The general phenomenon is true whether we are talking about a law, a pact among friends, an organizational norm, or a treaty between nations. This paper is not about designing a law where all but the desperate and the deviant will respect it. This model has no agent asymmetry: all players are identical. The paper's central result implies that even in the case where a law is tailor-made for a homogeneous population, it will not be fully adhered to by any member of the

[4]A partial contribution strategy is activity at less than 100% of the defined contribution. Notice the fundamental difference between occasional full non-compliance and continuous but minor deviations from full compliance: it is this latter, persistent shirking that needs explaining; the model will include the former as a punishment mechanism to increase the overall level of compliance.

community when monitoring is imperfect. Agent diversity, asymmetric monitoring costs, and variations in future payoffs may provide incentives for some agents to deviate more and others less.[5] But shirking will remain.

The paper proceeds as follows. Section 1 provides a brief background to the theory of compliance. Section 2 develops a general model of the compliance problem with a continuous action space. Section 3 establishes the unrealistic conditions sufficient for full compliance. Section 4 shows that under more plausible assumptions, full compliance is not possible. It also describes the conditions necessary to sustain any compliance in the face of inherent shirking. Section 5 discusses extensions of the model and concludes.

Background

To introduce the literature on compliance, we begin with a simple model. Later, in section 3, we will elaborate upon this model when the analysis requires more technical detail, but for now, consider a group of undifferentiated, symmetric agents who can coordinate their efforts to produce a non-rival, non-excludable public good. Here we have in mind a broad conception of a public good: a stable federation, reduction in nuclear weapons, establishment of free trade, cleaner air. An agreement specifies the actions expected of each participant in order to produce the public good optimally. No third party exists to enforce the agreement; it must be self-enforced. In each period agents choose how much to comply with the terms of the agreement. The overall benefit, or utility, that an agent receives each period is a function of its own compliance, the degree of compliance of all others, and any random interference. In this class of problems, agents are tempted to free ride, to let others make the sacrifices while the shirking agent reaps the rewards. In the formal model, we capture this temptation with a trio of assumptions that each agent's utility decreases as it complies more, increases as others comply more, and if everyone makes the same contribution, each agent prefers full

[5]For example, distributional issues are present when multiple equilibria exist and players are asymmetric (Krasner 1991, Morrow 1994, Weingast 1997, and Fearon 1998). Uncertainty about future payoffs makes flexibility advisable, either through exit clauses (Rosendorff and Milner 2001) or compromises on expectations (Downs & Rocke 1995, Downs, Rocke, & Barsoom 1996).

compliance by all to no compliance by any.

The most fundamental problem of this class is the Prisoner's Dilemma. In the one-shot version, two agents are limited to the binary choice of comply (fully) or defect. The sole predicted outcome to this problem is that both players defect. Expanding the field to multiple agents, in a version more closely associated with a public good provision problem, does not alter our prediction. Even repeating the game, whether for a few periods or many, does not improve the prediction if the game has a known end point. No one complies at all.

When the game is repeated indefinitely, however, compliance improves considerably. If the game is indefinitely repeated (or the players do not know when the game will end, and the discount factor is sufficiently high), then full cooperation can be maintained with trigger strategies, where some commonly-perceived information, correlated with agents' past play, cues agents about the play of others. Agents establish a threshold for this information variable; failure to meet it "triggers" a punishment phase. By the folk theorem we know that as long as the agents are patient enough, full compliance can be sustained with either a grim trigger strategy or the instigation of a finite punishment regime. It makes no difference what they employ, as in equilibrium no players deviate.[6]

We now add one element of complication to the model: noise. We keep all other dimensions of the model the same, that is, it is an indefinitely-repeated game played by multiple agents with a discrete choice space. However, it is now more difficult to infer agent action by the amount of the good provided, as random shocks intervene, distorting the translation of the group's effort into production, for better or worse. We now have a game with imperfect monitoring, as the trigger variable is only imperfectly correlated with agent behavior.

With imperfect monitoring, full compliance remains possible in this discrete choice world, under assumptions discussed below. However, the efficient equilibrium strategy is often based on a finite punishment mechanism and can be thought of as a generalization of tit-

[6]However, we may prefer the finite punishment regime if agents have cognitive constraints, to recover from agent error or experimentation. For example, recall from Axelrod's (1984) tournament, where players had to write their strategies down beforehand—a form of cognitive constraint—the forgiving strategy Tit-For-Tat outperformed Grim Trigger. See also Ostrom 1999.

for-tat. In the efficient equilibria, punishment regimes do occur, but they are triggered not by non-compliance, but due to the stochastic element (Abreu, Pearce, and Stacchetti 1986, 1990; Fudenberg & Maskin 1986; Fudenberg, Levine, and Maskin 1994). A finite punishment regime allows the cooperative regime to be restored. The term "punishment regime" is a misnomer: in equilibrium no one shirks; the punishment is meted out to insure all agents' future full compliance. Therefore we might more accurately call these "compliance maintenance" regimes.

At this point we should pause to consider the ability of a discrete choice model to explain non-attainment of full compliance. With only a binary choice of actions, partial compliance is not even a possibility. However, suppose that in addition to choosing between no compliance and full compliance, agents may choose among a finite set of partial compliance actions to allow the possibility of partial compliance equilibria (or might select mixed strategy equilibria). It is important to note that the existence of partial compliance equilibria does not rule out full compliance: by Fudenberg, Levine, and Maskin (1994)'s contribution to the folk theorem, we know that for any finite choice set, full compliance is an efficient equilibrium as long as two conditions hold: (1) if there exists a discount factor sufficiently high and (2) the signal imparts statistical identification of a deviator. Therefore, to explain partial compliance—that is, the absence of full compliance—we can apply what we might think of as the anti-folk theorem: if the discount factor is too low, full compliance cannot be sustained. It is possible to fix a low discount factor in advance to rule out the possibility of full compliance. However, such a result is undesirable: it has a feel of being forced, because to explain partial compliance and eliminate full compliance we must either construct impatient agents or insist that they are choosing inefficient equilibria!

Fudenberg, Levine, and Maskin's second condition alters the standard assumptions of imperfect monitoring. They model an asymmetric equilibrium, where players play slightly different strategies. This asymmetry means that in the long run, you can distinguish between any two configurations of discrete signals: the signal identifies the deviator. With this assumption, in the limit, players come arbitrarily close to full compliance. Their model

does not describe this paper, where the signal is public, common knowledge, and classically imperfect: it does not contain information that links outcomes to individual actions.[7]

To capture shirking, we need a continuous action space. Unfortunately, no folk theorem result extends beyond a world of discrete choice. The most general result is Green and Porter (1984), modeling cartel behavior, who characterize the optimal strategy for oligopolists, which they describe as monopolistic production punctuated by reversionary periods, "price wars" where cartel members produce at competitive levels.[8] These price wars occur when the noise term, a shock to consumer demand, drives price (the trigger mechanism) below a threshold. Green & Porter note (1984:93) that monopolistic levels of production might not be sustainable by a cartel. That is, they might not be able to sustain full compliance. The goal of Green & Porter is to characterize an optimal strategy generally, so they do not characterize the properties that make full compliance impossible.

This paper characterizes the compliance properties of any equilibrium in general collective action problems. First, for most conditions, *full compliance cannot be an equilibrium strategy*. Therefore it follows as an obvious corollary that full compliance cannot be an efficient equilibrium, since it cannot be an equilibrium at all. Note that if full compliance were an equilibrium, even if it were not the efficient equilibrium, its focality, its political appeal, and its legal enforceability might still recommend it. For political science these results are at least as important, and perhaps more important, than the specifics about how to implement the efficient equilibrium. While political environments contain adjudicatory institutions whose intervention might prevent spurious—and costly—punishment regimes, only a perfect monitor eliminates shirking altogether.

A second contribution of this paper is that it creates an environment for understanding what conditions affect levels of compliance. Specifically, with this model, we may manipulate

[7]Another related monitoring alternative is the literature on private monitoring, where players receive private signals about one another's actions. With private monitoring in repeated trade between two parties, where parties can choose between three actions (and any mixed strategy combination), Bhaskar and Van Damme (2002) show that full compliance is not supported for much of the parameter space.

[8]Downs and Rocke 1995 provides an excellent introduction of the Green & Porter model to political scientists.

the institutional setting to affect the probability that the punishment regime is triggered. I will return to this point in the conclusion.

The Model

The previous section referred to a basic model of public good provision problems. Here I summarize and complete the model's specification. Our question is whether—or under what conditions—full compliance can be maintained. We know by the folk theorems that full compliance is possible for discrete choice sets, at least under the conditions established by Fudenberg, Levine, and Maskin (1994). Therefore, this model begins with the continuous action space.

An agreement prescribes agent actions to provide the public good. At the start of each period t, all agents, $i, i \in \{1, \ldots, n\}$, choose a degree of compliance with the agreement, $c_{it}, c \in [0, 1]$, with $c_{it} = 1$ representing full compliance. I suppress t when doing so causes no confusion.[9] A nonatomic, continuously-distributed random variable, ω, captures stochastic uncertainty; it is independently drawn each period, unobserved by any agent, and has an expected value of zero: $E(\omega) = 0$.[10] Each agent's stage game utility is a function of its own action, the actions of others, and the stochastic element, so that $U_i = U_i(c_i, c_{-i}, \omega)$. Agent utility decreases in its own level of compliance, $\frac{\partial U_i}{\partial c_i} < 0, \forall i$, but increases in the level of compliance of all other players, $\frac{\partial U_i}{\partial c_j} > 0, \forall i, \forall j \neq i$. For simplicity, let the vector $\vec{c}$ represent the symmetric action taken by all participants, where $\vec{c} = 1$ represents full compliance by all. In particular, $U_i(\vec{c} = 1) > U_i(\vec{c} = 0)$; full compliance by all is strictly preferable to complete non-compliance. Agent impatience is represented by a non-zero discount rate, $\delta \in (0, 1)$. Agents are risk neutral and maximize expected utility $E[\sum_{t=0}^{\infty} \delta^t U_{it}(c_i, c_{-i}, \omega)]$.

[9] Scholars in other disciplines have loosened the discrete time restriction. See for example Boyd 1988, Huberman & Glance 1993, Weesie & Wippler 1987.

[10] Since the stochastic term can assume positive or negative values, the model can capture unexpected benefits as well as unexpected losses. Compare this model to the decentralized decision-making model of Bendor and Mookherjee (1987), where uncertainty enters as a probability that a player's effort will not successfully translate into a contribution.

Trigger Variables and Mechanisms

Most groups faced with the provision of public goods employ some institution to induce cooperation (Ostrom 1990). Here, I represent these institutions simply as a trigger mechanism, where agents' strategies are contingent on a signal generated by their collective levels of compliance and a stochastic element. If the trigger variable—the signal—falls below a threshold, τ, players refrain from complying for a finite number of periods, T.[11] In order to be effective, the trigger mechanism, including the signal, must be common knowledge.

The realization of the trigger variable is a function F of each agent's level of compliance and the noise term, so $F_t = F(\vec{c}_t, \omega_t)$.[12] I assume that F is symmetric relative to the agent contributing, where $\frac{\partial F}{\partial c_i} = \frac{\partial F}{\partial c_j} \quad \forall i$ and j. For example, in a public goods problem the function F could equal the sum of the contributions plus the noise term: $\sum c_i + \omega$, or in Cournot competition, the signal could be the market price, as in Green and Porter (1984) and Downs and Rocke (1995). However, notice that the signal does not appear in the agent's utility function. We need not assume that agents get any utility from the signal; it may simply be an indicator they use to determine behavior. For example, if agents are participating in a treaty to ban nuclear testing, a signal may be the measurement of tremors in the earth to detect underground tests. While these tremors do not affect the utility of the participating nations, if tremors are detected, nations alter their behavior.

Given the restrictions on information, agents are restricted to signal-contingent strategies. A signal-contingent strategy for i is an infinite sequence of actions $s_{it} = \{s_{i0}, s_{i1}, \ldots\}$ where s_{i0} is i's initial level of compliance c_{i0}, and thereafter compliance at time $t+1$ is a function of the value of all past signals, so $c_{i,t+1} = s_{i,t+1}(F_0, \ldots, F_t)$. To prove the results of this paper, I can restrict attention to those strategies that involve finite punishment with maximal punishment in each period of the punishment regime. The analysis trivially extends to

[11] It is important to keep in mind that the punishment regime is not the same as withdrawal from the group. Players still "participate," but they do not comply. Note that the results carry through entirely if agents are monitored and punished individually as long as there is imperfect monitoring, although specific solutions would shift slightly.

[12] For simplicity I have used the same notation for the stochastic element here. One might prefer to distinguish the noise term in the utility function from the one generating the signal.

infinite punishment regimes. Maximal punishment in any period of the punishment regime elicits the highest degree of compliance. Therefore we can restrict attention to this case.

To characterize this class of strategies, define t to be a *normal* period if any of the following three conditions hold: (1) $t = 0$, (2) $t-1$ is normal and the signal at $t-1$ exceeded the threshold: $F_{t-1} > \tau$, or (3) the agents are emerging from a punishment regime: $t - T$ was normal (by condition 2) and $F_{t-T} < \tau$. If none of these conditions hold, define t to be a *reversionary* period of the punishment regime.

Any strategy must take the following form:

$$c_{it} = \begin{cases} a_i & \text{if } t \text{ is normal} \\ 0 & \text{if } t \text{ is reversionary.} \end{cases}$$

where a_i is the degree of compliance by agent i in a normal period, $a_i \in [0, 1]$. It will be useful to define $\theta = 1 - a_i$, where θ is a measure of shirking. Let X_i represent the utility received in each period of the punishment regime. I assume that $U_i(\vec{a}, \omega) \geq X_i$ for any symmetric a_i: agents prefer any degree of compliance made by all agents to their payoff in the punishment regime. (Hereafter I drop the subscript for X, since the agents are symmetric.) This paper will first characterize the conditions necessary to sustain full compliance, where $a_i = 1$ and $\theta = 0$, and then will examine the conditions that sustain partial compliance, where $a_i > 0$ but $\theta > 0$.

A Nash equilibrium is a strategy profile $(s_1^*, \ldots, s_n^*)$ which satisfies

$$E_{s_1^*,\ldots,s_i^*,\ldots,s_n^*}\left[\sum_{t=0}^{\infty} \delta^t U_{it}\left(c_{it}^*, c_{-i,t}^*, \omega\right)\right] \geq E_{s_1^*,\ldots,s_i,\ldots,s_n^*}\left[\sum_{t=0}^{\infty} \delta^t U_{it}\left(c_{it}, c_{-i,t}^*, \omega\right)\right] \qquad (1)$$

for all agents and feasible strategies s_i.[13]

The Value Function

Agents maximize V_i, the present discounted value of their utility. To solve for V_i, I construct the following two-state T-stage Markov dynamic programming problem. Let a^* be a equi-

[13] I focus on the Nash equilibrium as I am considering agents with foresight. It is also possible to take an evolutionary perspective, which is of particular value with games of smaller stakes and many spatially organized actors. A good introductory text is Gintis 2000; for representative scholarship, see Bendor & Swistak 1997.

librium candidate for play in normal periods. The present discounted value for agent i of $a_i|a^*_{-i}$ is defined as follows:

$$V_i(a_i|a^*_{-i}) = U_i(a_i, a^*_{-i}, \omega) + (1-p)\delta V_i(a_i|a^*_{-i}) + p\left(\Delta^T X + \delta^{T+1} V(a_i|a^*_{-i})\right) \tag{2}$$

where $\Delta^T = \delta + \delta^2 + \cdots + \delta^T$, the sum of all discount parameters to the T^{th} period and p is the probability that the signal falls short of the threshold, triggering a punishment regime: $p = \text{prob}\left(F(\vec{a}, \omega) < \tau\right)$. Note that p is not a parameter, but is a function of the agents' actions. The value function captures both the costs and benefits of shirking (the temptation to contribute less than $a_i = 1$ in a cooperative phase). The first term on the right-hand side is the per-period utility for a given level of contribution, while together the second and third terms are the continuation value, where the second term is the continuation value to agent i of successful coordination (when $F_t(\vec{a}, \omega) > \tau$), and the third term the remaining cases, when $F_t(\vec{a}, \omega) < \tau$ and the group enters a punishment regime.

The vector of actions a^* is an equilibrium in the per-period cooperative phase if and only if

$$V_i(a^*_i, a^*_{-i}) \geq V_i(a_i, a^*_{-i}) \quad \text{for all} \;\; a_i \in [0, 1].$$

For example, if V_i is concave in a_i for all i (where $\frac{\partial^2 V_i(a_i, a^*_{-i})}{\partial a_i^2} \leq 0$), then a necessary and sufficient condition for equilibrium is $\frac{\partial V_i(a_i, a^*_{-i})}{\partial a_i} = 0$ at $a_i = a^*_i$: here agents maximize their own utility given the play of others.

Each player selects its optimal strategy by making a calculation about what it expects to gain by deviating from full compliance (the first term of Eqn. 2). It compares this benefit of deviation to the cost of deviation, which is the increased probability that F falls below the threshold τ. Deviating offers short-term gain, but comes at a price of increasing the likelihood that the group falls out of the cooperative equilibrium and into an unproductive punishment regime. The agent deviates until the marginal benefit of deviation equals the marginal cost.

To find the value of a^*_i that maximizes the value function, we solve as follows. First,

letting V and U represent $V_i(a_i, a_{-i}, \omega)$ and $U_i(a_i, a_{-i}, \omega)$, respectively, Eqn. 2 simplifies to:

$$V = \frac{U + p\Delta^T X}{(1-\delta) + p(\delta - \delta^{T+1})}. \tag{3}$$

Eqn. 3 provides a generic value function for this class of problems.

From Eqn. 3 we take the derivative w.r.t. a_i:

$$V' = \frac{\left((1-\delta) + p(\delta - \delta^{T+1})\right)\left(\frac{\partial U}{\partial a_i} + \frac{\partial p}{\partial a_i}\Delta^T X\right) - (U + p\Delta^T X)\left(\frac{\partial p}{\partial a_i}(\delta - \delta^{T+1})\right)}{\left((1-\delta) + p(\delta - \delta^{T+1})\right)^2} \tag{4}$$

An equilibrium satisfies $V' = 0$, so it suffices to drop the denominator, which is strictly positive:

$$0 = \frac{\partial U}{\partial a_i}\left((1-\delta) + p(\delta - \delta^{T+1})\right) - \frac{\partial p}{\partial a_i}\left((U - X)(\delta - \delta^{T+1})\right) \tag{5}$$

Eqn. 5 specifies a necessary condition for a level of compliance that maximizes the value function. The first term specifies the marginal benefit of complying and the second, the marginal cost. In equilibrium these terms must be equal.

Conditions Necessary to Sustain Full Compliance

Complete Information

To derive intuition about strategies, we first consider the full information case, where we suppress the random variable ω, so the total production plainly reveals whether or not any agent shirked. Each agent chooses an action in normal periods $a_i \in [0, 1]$. Each agent's payoff is a function of the contributions of all agents and its own opportunism, if any, so that the single period utility function is $U = U(a_i, a_{-i})$.

While many equilibria exist in the repeated game setting, we concentrate on the full compliance equilibrium, where $a_i = 1$. As long as the discount rate is sufficiently low (that is, the discount factor δ, must be sufficiently high), full compliance can be sustained with a grim trigger punishment mechanism. Strategies are contingent upon the play in the previous round: each agent cooperates until one defects, at which point the other agents pull the "grim

trigger," and defect forever.[14] To place this special case in our general formulation, we need only assume that F is a deterministic function: the signal contains no random component. For example, the signal might be the sum of the compliance levels. Therefore, we can set the threshold, τ, equal to the full compliance signal $F(\vec{1})$.

Proposition 1: *With complete information, full compliance can be achieved with a trigger strategy in which deviations are punished forever by setting the threshold, τ, equal to the signal generated by full compliance by each member.*

$$\begin{aligned} c_{i,0} &= a_i = 1 \\ c_{it} &= \begin{cases} a_i = 1 & \text{if} \quad F_{t-1}(\vec{a}) = F(\vec{1}) \\ 0 & \text{else.} \end{cases} \end{aligned}$$

PROOF: It suffices to show that no player can benefit from shirking. The player compares its expected utility of full compliance with the expected utility from shirking, choosing the action that gives it the higher expected payoff. Without any information uncertainty, any shirking is revealed. Therefore, the threshold, τ, is set such that it tolerates no deviation from full compliance. By assumption, $U_i(\vec{a} = 1) > U_i(\vec{a} = 0)$ and $U_i(a) \geq X$, for all a, so no player would prefer to deviate, knowing that to do so would certainly trigger the punishment regime, for a sufficiently high discount rate.[15]

While the grim trigger strategy is theoretically efficient, giving the most bang for buck in terms of ensuring compliance, it is severe and unforgiving. It might not be renegotiation proof, a subject I will return to briefly in the conclusion. It is also irrelevant from a practical perspective: no player would choose to remain forever in a non-complying, non-productive group, if it could withdraw and wash its hands of the matter. In practice, we would expect that the grim trigger strategy would become a strategy of dissolution.[16]

Another strategy option exists: full compliance can also be sustained with a finite punishment strategy, where players respond to less-than-full compliance with non-compliance

[14] This result is similar in feel to the folk theorem. However, no folk theorem results are available for continuous action spaces with imperfect monitoring.

[15] See the proof for Proposition 2 in the Appendix.

[16] Extensions of this research will include provisions for exit options.

for a specified number of periods, rather than forever, as in the case with the grim trigger.

Proposition 2: *With complete information, the full compliance equilibrium can be sustained with a finite punishment regime, defined as follows:*

$$\begin{aligned} c_{i,0} &= a_i = 1 \\ c_{it} &= \begin{cases} a_i = 1 & \text{if } t \text{ is normal} \\ 0 & \text{if } t \text{ is reversionary.} \end{cases} \end{aligned}$$

PROOF: See appendix.

The proof generates the following comparative statics: T, the minimum number of punishment periods, will vary inversely with $U_i(\vec{a} = 1)$ and δ, the benefit of full compliance and the discount factor, respectively.[17] Therefore, as the players become increasingly patient, the minimal number of punishment periods necessary to sustain full compliance declines. Furthermore, if the signal is related to utility, and the translation of compliance into return is made more efficient (for different functions F), players are increasingly willing to comply, so fewer punishment periods are needed to sustain full compliance. These results continue to hold in the more general cases below.

If a grim trigger strategy is in practice equivalent to permanent withdrawal, the finite punishment regime is the strategy that makes forgiveness and regeneration of cooperation possible. In the finite punishment regime, no agent contributes, so the group is not productive, but the nominal association persists, ready to be revived at a later date. It is the most primitive of institutional mechanisms available to coerce compliance. Implementing the punishment mechanism sacrifices the benefits of cooperation for some time, but it does not entail further costs, such as dipping into the reserved benefit (a_i). The next most primitive method to induce compliance is interagent conflict, a means that does require personal sacrifice beyond the loss of opportunity. These measures are implemented after undesirable outcomes are felt, but in equilibrium they operate as ex ante deterrents, and are only used when environmental circumstances generate unfavorable levels of productivity, a condition without relevance unless we include imperfect monitoring. Here, no agent shirks.

[17] See Eqns 10 & 12 in the Appendix.

Incomplete Information and Full Compliance

The section above demonstrates that with full information, a full compliance equilibrium can be sustained with a trigger strategy. However, even with incomplete information certain conditions support full compliance in equilibrium, where $\theta = 0$, but these conditions demand implausible assumptions about the cost and benefit functions, making full compliance highly improbable.

Proposition 3: *If the probability of punishment* and *the benefit of deviating are linear in the level of compliance, then any symmetric stable equilibrium requires either full compliance or no compliance. No stable partial compliance equilibria exist.*

PROOF: For a formal proof, see appendix. The agents comply at the level where the marginal cost of deviating equals the marginal benefit of deviation. If the cost and benefit curves are linear in the amount of compliance, then the marginal cost and benefit curves are level at a constant, giving us just three possibilities: (1) the marginal cost is always greater than the marginal benefit, for all levels of compliance, (2) the marginal cost is equal to the marginal benefit, or (3) the marginal cost is less than the marginal benefit of deviation. If the first is true, then no agent will be tempted to deviate from full compliance.[18] If the second is true, while equilibria with slippage may exist, they will not be stable. In the final case, agents will deviate fully, and no public good is provided.

In essence, linearity returns us to the discrete case, and so we should not be surprised that full compliance can be sustained. However, linearity was just a mathematical convenience: we can reconstruct Proposition 3 to include non-linear functions. Recall that the measure of shirking is θ. Consider the benefit and cost functions of shirking, $B(\theta)$ and $C(\theta)$, where $B(0) = C(0) = 0$.

[18] And, since the severity of the punishment is manipulable by the agents, they will probably adjust the marginal cost curve downward so that it just surpasses the marginal benefit curve.

Corollary 3.1: *If $B(\theta) \leq C(\theta)$, $\forall\, \theta$, then full compliance is supportable in equilibrium.*

Corollary 3.2: *If $C(\theta) \leq B(\theta)$, $\forall\, \theta$, then no compliance is supportable in equilibrium.*

Proof: Follows intuitively.

Proposition 3 and its corollaries indicate that full compliance is possible. By Proposition 3 we know that the only stable equilibria if both the cost and benefit functions are linear is full compliance or none. The two corollaries suggest that linearity is not important: there are conditions that would support full compliance with non-linear functions. However, the next section will establish the difficulty in satisfying these conditions, and therefore the implausibility of full compliance. It will also generate conditions that sustain some level of partial compliance: unlike the linear case, where the only two stable possibilities are full compliance or none, with nonlinear functions we can sustain partial compliance.

Modeling Inherent Shirking

The Logic

In the previous section, I suggested that the conditions necessary to sustain full compliance were unlikely to be met by any real-world circumstances. We should satisfy ourselves that these circumstances are indeed unlikely by considering further what these results imply. Perhaps the best way to think about the unrealistic limitations of these propositions in political contexts is by discussing what they exclude. Consider the following four assumptions.

A1: If everyone is complying fully, the marginal benefit of shirking is high.

A2: However, thereafter, the marginal benefit of shirking decreases as the amount of shirking increases.

A3: If everyone else is complying fully, the initial probability that a small degree of shirking triggers the punishment mechanism is very low.

A4: The probability of triggering the punishment mechanism increases as an agent moves from full compliance.

If either A2, diminishing marginal utility or A4, increasing marginal cost, holds, Proposition 3 is negated. Diminishing marginal utility is a standard assumption. In the context of a public goods provision problem, consider the value to the agent of diminishing his effort by a single generic "unit." Diminishing marginal utility implies that the value of decreasing effort by this unit is higher if overall the agent is complying very much than if the agent is already shirking considerably: the first chemical weapon a nation builds covertly is worth more to it than its 100th (A2) and is also harder to detect (A3); as a member of a protesting crowd confronting police, moving from the first row to the second is worth more to the shirker than trading places at the back. The instances where a game of compliance will not exhibit diminishing marginal benefits from shirking are rare.

Assumption A2 generates a concave benefit function from shirking, while assumption A4 generates a non-linear cost function from shirking. Recall that the cost of shirking is a calculation of the probability of getting caught (when the value of the good falls below the threshold) multiplied by the loss incurred when the group enters a non-cooperative punishment regime. Because of environmental uncertainty, the threshold to trigger the punishment regime contains a little flexibility. For high levels of compliance, where ($\vec{a} \simeq 1$), the marginal effect of shirking a bit more is unlikely to trigger the punishment regime. Equivalently, when agents are hardly complying ($\vec{a} \simeq 0$), the effect of marginally increasing the level of compliance is likely to be negligible, as the group is almost certain to enter the punishment regime anyway. However, for some intermediate levels of compliance, small deviations may have a large effect on the probability of punishment, acting as the proverbial straw that broke the camel's back. Instances that violate both conditions are rare. The presence of either of the above conditions is sufficient to nullify Proposition 3 because it introduces non-linearity in the cost or benefit of compliance.

Corollaries 3.1 and 3.2 did not require linearity to sustain full compliance. As long as the marginal cost of shirking is always greater than the marginal benefit, full compliance is

supportable. However, while these two corollaries embrace diminishing marginal utility and increasing marginal costs (A2 & A4), they violate A1 and A3.

Assumption A1 states that if all other agents are in full compliance, the marginal benefit to an agent of shirking is very high. In order to resist the temptation to shirk, the marginal cost of shirking must also be very high. But A3 asserts the opposite: if all others are complying fully, the probability that a small amount of shirking would trigger a punishment regime nears zero. While it is possible that the pain of punishment is so severe that even with this tiny probability, the marginal cost of shirking exceeds the marginal benefit, such an assumption would need to be justified explicitly within the context of the specific circumstance, as it would hardly be a trivial detail of the model. Therefore, the contradiction between A1 and A3 makes it highly unlikely that the conditions necessary to sustain Corollaries 3.1 and 3.2 will be met.

To help clarify the previous discussion, it is useful to consider the most realistic case: concave marginal utility and an S-shaped probability of triggering the punishment regime. In this case, all four assumptions are satisfied. The concave utility function implies that the marginal benefit of shirking is highest at full compliance [A1] and that thereafter it decreases [A2]. The S-shaped probability of triggering the punishment regime as a function of shirking implies that small degrees of shirking are unlikely to trigger the punishment regime when all comply fully [A3]. The intermediate steep slope implies that the marginal cost is increasing as you move from full compliance to small amounts of shirking [A4]. The implausibility of full compliance is overdetermined.

Modeling Partial Compliance

We now formally express the conditions in section 5.1 and give conditions to support partial compliance in equilibrium. To guarantee the possibility of any compliance, the marginal benefits and marginal costs of shirking must be equal within the interior of the set of possible actions (otherwise, no compliance is the only equilibrium). Informally, the marginal benefit of shirking must start out above the marginal cost at $a_i = 1$, and, as the level of compli-

ance drops to zero, it must fall below the marginal cost. The point of intersection of the marginal benefit and marginal cost curves will identify the equilibrium level of compliance. The mathematical argument below characterizes the curvature conditions to sustain partial compliance.

We return to Eqn. 4 to find the value of a_i that maximizes the value function. We are interested in equilibria where V_i' is maximized for values of a_i less than one but greater than zero. The first condition guarantees slippage, the second guarantees some minimal compliance. We make the following assumption: V_i is quasi-concave in a_i.[19] A stronger but sufficient condition, for the existence of an equilibrium, is that V_i' is concave, which will be true if the second derivative of the value function (Eqn. 3) is negative, which reduces to:

$$\begin{aligned}&[(1-\delta)+p\Delta]^2[U''((1-\delta)+p\Delta)-p''(U-X)\Delta]\\&\quad-2[(1-\delta)+p\Delta]p'\Delta[U'(1-\delta+p\Delta)-p'(U-X)\Delta] \;<\; 0. \qquad (6)\end{aligned}$$

Proposition 4: *Given that the value function, V_i, is quasi-concave, there exists a pure strategy equilibrium to this game.*

PROOF: See appendix.

By Proposition 4, we know that an equilibrium exists to the game in which players comply at some level $a_i \geq 0$. The following three propositions establish two sufficient conditions to guarantee that no full-compliance equilibrium exists: either the benefits of shirking must be large if everyone else is complying, or the likelihood of triggering the punishment regime must be small if one player shirks and the rest comply. Players shirk if *either* condition is present; players will "agree" upon some level of deviance, and no equilibria will include full compliance in the strategy set. While it is highly improbable that either the utility function

[19] It is not difficult to derive sufficient conditions for quasi-concavity, but the algebra becomes messy. However, the conditions reduce to the following insights: (1) the change in marginal utility has to be large relative to the change in marginal probability of being caught or (2) the marginal probability of being caught must be small.

or the probability of punishment is linear in the amount a player contributes (as discussed above) Prop'ns 5 & 6 each isolate the two sufficient conditions by restricting one of the two functions to linearity. We first consider how concave benefit functions imply shirking.

Proposition 5: [Concave benefits, linear probability of punishment] *Suppose* $\frac{\partial p}{\partial a_i} = -\beta$. *If (i)* $\frac{\partial^2 U_i}{\partial a_i^2} < 0$, *(ii)* $\frac{\partial U_i}{\partial a_i} \to 0$ *when* $\vec{a} \to 0$ *and (iii)* $\frac{\partial U_i}{\partial a_i} < -\beta \left(\frac{\delta(U_i(\vec{a}=1)-X)}{(1-\delta)} \right)$ *at* $\vec{a} = 1$, *then the equilibrium will not support full compliance, but will support some compliance.*

PROOF: See appendix.

The conditions in Proposition 5 are straightforward to interpret. The first condition guarantees that utility is maximized at some point.[20] The second condition says that there are diminishing marginal returns from shirking; complying a little bit doesn't hurt you all that much for very low levels of compliance, thereby ensuring some compliance in equilibrium. Under the third condition, if all others are complying fully, it becomes very tempting for an agent to shirk. If the marginal utility of shirking is small, particularly relative to the potential punishment, then as the players become more patient, full compliance becomes more likely. However, the greater the marginal utility of shirking a small amount, the less likely full compliance becomes. If these conditions on the utility function hold, and the probability of punishment is linear in the level of compliance, then no equilibrium satisfies full compliance.

Consider how the general results apply to a common specific functional form. Suppose that an agent's utility is a function of the sum of all agents' compliance (because of the public good properties) and of the degree to which the agent shirks (a private good), can be expressed as a standard Cobb-Douglas utility function:

$$U_i(a_i) = \left(\sum_{i=1}^{n} a \right)^{\alpha} (1 - a_i)^{\beta},$$

[20] I'm assuming that the utility function is concave, although it need not be to generate a quasi-concave value function.

with $\alpha, \beta \in (0,1)$. The partial derivative with respect to a_i,

$$\frac{\partial U_i}{\partial a_i} = \left[\alpha \left(\sum_{i=1}^{n} a\right)^{\alpha-1}\right](1-a_i)^{\beta} - \left(\sum_{i=1}^{n} a\right)^{\alpha} \beta\,(1-a_i)^{\beta-1},$$

converges to $-\infty$ at $\vec{a} = 1$. For this standard functional form of per-period utility, full compliance cannot be an equilibrium for any $\delta < 1$. The intuition that we get from the folk theorem—that as the discount rate goes to 1, full compliance can be sustained—does not hold.[21] Alternatively, if we chose a functional form such that the marginal utility at $\vec{a} = 1$ was negative and *finite*, then there would exist a threshold discount rate, δ^*, such that if $\delta > \delta^*$, then full compliance would be possible. It is possible to select a particular functional form where the folk theorem logic applies, but the logic does not apply generally.

Proposition 6 inverts Proposition 5 to show how a non-linear probability of punishment guarantees positive levels of shirking. Although we've already underscored the conventionality of assuming a strictly concave benefits function, shirking can occur even when the function is linear. Letting the benefits function be linear in the level of compliance, Proposition 6 establishes a sufficient condition for shirking that depends upon the change in likelihood of triggering the punishment regime. The probability of punishment must approach zero as the participants reach near-full compliance, and for some level of compliance the probability of triggering the punishment regime for any marginal deviation must be high. The function itself must be non-linear, but may be convex or take an S-shape.

Proposition 6: [Linear benefits function, non-linear probability of punishment] *Suppose* $\frac{\partial U_i}{\partial a_i} = -\alpha$. *If an equilibrium exists, it does not support a strategy of full compliance if (i)* $\frac{\partial p}{\partial a_i} \to 0$ *as* $\vec{a} \to 1$. *Furthermore, if (ii)* $\exists\ \bar{a}$ *for which* $\frac{\partial p}{\partial a_i} < \frac{-\alpha(1-\delta^{T+1})}{(U_i(\bar{a})-X)(\delta-\delta^{T+1})}$ *at* $\vec{a} = \bar{a}$, *then the equilibrium supports some compliance.*

PROOF: See appendix.

[21] At $\delta = 1$, we are comparing two infinities. A numerical example with a different utility function is developed by Kreps (1990:517-521). Here a reader can see the effect of increasing the number of punishment periods and changing the sensitivity of the threshold, as well as the inability, in his example, to obtain full compliance.

In Proposition 6, it is the first condition that guarantees slippage; if there is very little chance that one player will trigger the punishment regime if everyone else is complying fully, then the agent will have little incentive to resist the urge to shirk. Furthermore, the proposition states that in order to guarantee some compliance, there must be some point where an increase in one player's level of shirking becomes very likely to trigger the punishment regime. If the second condition is not met, the equilibrium strategy is full deviation and the group probably would not exist.

Propositions 5 & 6 relaxed the linearity assumption on $U_i(a)$ and $p(a)$ in turn. Proposition 7 relaxes both assumptions, showing that weaker of the conditions in Prop'ns 5 & 6 is sufficient to guarantee non-existence of full compliance equilibria.

Proposition 7: [Concave utility, non-linear probability] *If the probability of punishment, $p(a_i)$, is non-linear in a_i and the utility function, U_i, is concave in a_i, then no full compliance equilibrium exists provided (i)* $\frac{\partial U_i}{\partial a_i} < \frac{\partial p}{\partial a_i}\left(\frac{\delta(U_i(\vec{a}=1)-X)}{(1-\delta)}\right)$ *at* $\vec{a}=1$. *Furthermore, if (ii)* $\exists\ \bar{a}$ *s.t.* $\frac{\partial p}{\partial a_i} < \frac{\frac{\partial U_i}{\partial a_i}(1-\delta^{T+1})}{(U_i(\bar{a})-X)(\delta-\delta^{T+1})}$ *at* $\vec{a}=\bar{a}$, *then the equilibrium will support some compliance.*

PROOF: Follows directly from Propositions 5 & 6.

Taken together, these propositions demonstrate why full compliance equilibria are not supportable in most political contexts. To support full compliance, an agent must have a high probability of getting caught from shirking and get little marginal benefit. Both suppositions contradict logic and standard assumptions. The propositions also establish minimal conditions for partial compliance; if these conditions are not met, then the unique equilibrium is $\vec{a}=0$: no compliance.

Implicit in Propositions 5, 6, & 7 are a variety of comparative statics results on the equilibrium level of compliance that are easy to derive and empirically testable. The faster the decrease in the marginal benefit of shirking, the less shirking we will see. The longer the probability of triggering punishment stays low, the more shirking we will see. Given a particular context, such as those considered by Green & Porter or Downs & Rocke, it is

a straightforward exercise to show how levels of compliance vary with parameters of their models.

Discussion

This paper has shown theoretically what we have come to know empirically: for symmetric agents with a public, imperfect signal, full compliance is rarely possible in collective action problems and public goods games, such as international cooperation, many aspects of federalism, and collusion. Theoretical findings of full compliance equilibria rely on the simplifying assumptions of perfect monitoring or discrete action spaces that are unlikely to be met in reality. When compliance is possible, we are going to see single-period realizations that are neither "defect" nor "cooperate," but somewhere in between. We will never see (under the real-life conditions described here) full compliance. Therefore, some shirking should be the norm in games of compliance because any equilibrium includes deviation. Even when the group seems cooperative (that is, when not in a punishment regime), all agents shirk to a degree.

Notice that official expectations are not compromised. The agreement is drawn for full compliance although no one complies fully. Equilibrium behavior appears to be a "reasonable effort" and includes flexibility to accommodate moderate shirking from full compliance.[22] From time to time, and due to random bad shocks (and not further non-compliance on the part of the agents), the performance of the group falls to intolerable levels. With the primitive institutional mechanism we have constructed, all will occasionally deviate fully, for some finite period. This punishment regime is necessary to elicit all (partial) compliance.

Some analyses of specific functional forms provide an intuition about comparative statics (Green & Porter 1984, Kreps 1990, Downs & Rocke 1995). For example, if we assume that increasing the number of participants decreases each one's effect on the trigger variable, then overall compliance falls. In Green & Porter, price is the trigger variable, and with more participants in the cartel, each individual's action is less likely, on the margin, to change the

[22]Although it is beyond the scope of this paper, this result has implications for models of judicial review.

price, so cartel members shirk more, by producing more, as the size of the cartel grows. A second notable result is that as punishment becomes more severe, more compliance can be extracted, to a point: it is exactly this result that Downs, Rocke, and Barsoom 1996 turn to in identifying non-compliance as an enforcement problem. Other specific functional forms that fall within the general specifications in this paper may allow us to calibrate the extent of non-compliance for specific circumstances.

All of these analyses share the same weakness: multiplicity of equilibria. While for specific functional forms (and in some of the discrete choice analysis, eg. Abreu, Pearce, & Stachetti), closed form solutions generate efficient equilibria with optimal behavior, none of these analyses can guarantee that in reality the agents will settle upon the efficient equilibrium. This paper gives us a different kind of certainty: it tells us what will *not* happen: full compliance.

The result has important implications for legal and institutional design. Unless we recognize that full compliance is impossible, we will have unreasonable expectations that lead us to design inappropriate or inefficient rules. In particular, we may neglect the design or importance of institutions that effectively punish moderate shirking. If we assume that compliance could be full, then we may not seek institutional means to increase compliance, but if compliance is not full, then we may search for more innovative institutional solutions. For example, we may be prodded to investigate how institutions might work in tandem, and in complementing one another, increase overall compliance. Dispute resolution also becomes an important feature of institutional design when shirking is inherent.

Notice also that the punishment regime is a costly mechanism for keeping all players honest. We would prefer to have milder corrective mechanisms that avoid the sulky periodic retreat forecast by the punishment regime. Here enters a second role for institutional analysts: carefully designed institutions that combine fragmentation, monitoring and adjudication fulfill the function of the punishment regime without the high cost because they set up incentives to ensure compliance. A contrast with the Green and Porter setting is illustrative. In their model, "compliance" refers to the maintenance of a cartel—an illegal

activity—therefore, no legal remedy is available when the whims of consumer demand drive price below the threshold, and a price war must result. The firms cannot appeal to a court or other mediating institutions to get verification that no one deviated. In contrast, if air quality in the western United States becomes markedly worse, it is possible that the states could appeal to federal court only to find that an el niño was responsible for a six month inverted weather pattern and that in fact no shirking existed. The extent to which adjudicating institutions can prevent unnecessary punishment regimes varies. It is clearly higher within a political community than it is within the international arena, where adjudicatory institutions are not as readily available. The possibility and implications of such institutional interventions are addressed more fully in a companion paper (Bednar 2004). As long as the institutions fulfill their functional requirements, they may be tailored to meet local traditions for a greater chance of legitimacy.

A final consideration, usually ignored, is the necessity of establishing common knowledge to render trigger mechanisms feasible. In economic contexts, trigger mechanisms come relatively easily through the transparency of prices. However, in political contexts, the mechanism is often laden with subjective ambiguity. One role of institutions may be to establish the commonality of perception—through public revelation of information, through establishment of focal points—necessary to make a trigger mechanism operable.[23]

This paper has shown that the question is not *whether* but *how much* agents will shirk. We cannot eliminate shirking. The role for institutional analysts is to design institutions to *manage* this shirking (and its effects). Unless institutions are perfect monitors, they won't eliminate all shirking. But they might reduce or eliminate the need for costly punishment regimes, improving the gains from cooperation. And they may reduce the consequence of the shirking that remains.

[23]See Weingast's (1997:261) discussion of the role played by constitutions and elite pacts to facilitate coordination by declaring common expectations.

Appendix

Proof of Proposition 2: The infinite game full compliance equilibrium payoff and the payoff from deviations can be written as:

$$\begin{aligned} C^{\infty} &= U_i(1, 1_{-i}) \left(\frac{1}{1-\delta} \right) \\ D^{\infty} &= X \left(\frac{1}{1-\delta} \right) \end{aligned}$$

respectively.

In the finite punishment game, payoffs in the next T periods from cooperating and deviating are, respectively:

$$\begin{aligned} C^{T} &= U_i(1, 1_{-i}) + \sum_{t=1}^{T} \delta^t U_i(1, 1_{-i}) \\ D^{T} &= U_i(0, 1_{-i}) + \sum_{t=1}^{T} \delta^t X \end{aligned}$$

If $C^{\infty} > D^{\infty}$, then $\exists\, T^*$ s.t. $C^{T^*} > D^{T^*}$.

To derive comparative statics, we need to be more specific, and find the minimal T^* such that $C^T \geq D^T$. For simplicity, let U(C) and U(D) represent the utility from cooperating and deviating, respectively.

$$\begin{aligned} U_i(C) + \sum_{t=1}^{T} \delta^t U_i(C) &\geq U_i(D) + \sum_{t=1}^{T} \delta^t X \\ \sum_{t=1}^{T} \delta^t U_i(C) - \sum_{t=1}^{T} \delta^t X &\geq U_i(D) - U_i(C) \\ (U_i(C) - X) \sum_{t=1}^{T} \delta^t &\geq U_i(D) - U_i(C) \\ \sum_{t=1}^{T} \delta^t &\geq \frac{U_i(D) - U_i(C)}{U_i(C) - X} \\ \frac{\delta - \delta^{T+1}}{1-\delta} &\geq \frac{U_i(D) - U_i(C)}{U_i(C) - X} \\ \delta - \delta^{T+1} &\geq (1-\delta) \left[\frac{U_i(D) - U_i(C)}{U_i(C) - X} \right] \\ \delta^{T+1} &\leq \delta - (1-\delta) \left[\frac{U_i(D) - U_i(C)}{U_i(C) - X} \right] \end{aligned} \tag{7}$$

For simplicity, let $R = \delta - (1-\delta)\left[\frac{U_i(D)-U_i(C)}{U_i(C)-X}\right]$. Then:

$$\begin{aligned} \delta^{T+1} &\leq R \\ (T+1)ln\delta &\leq lnR \\ Tln\delta + ln\delta &\leq lnR \\ Tln\delta &\leq lnR - ln\delta \quad (8) \\ T^* &\geq \frac{lnR - ln\delta}{ln\delta} \quad (9) \end{aligned}$$

We now have a representation of the minimal number of punishment periods, T^*, which we can use to generate comparative statics regarding $U(C)$ and δ.

We use $\frac{\partial R}{\partial U(C)}$ to find $\frac{\partial T}{\partial U(C)}$. Using Eqn. 7:

$$\frac{\partial R}{\partial U(C)} = \frac{[U(C)-X](1-\delta)+(1-\delta)[U(D)-U(C)]}{[U(C)-X]^2}$$

Since we need only to sign the derivative, I suppress the denominator, as it is positive.

$$\begin{aligned} \frac{\partial R}{\partial U(C)} &= (1-\delta)[U(C)-X+U(D)-U(C)] \\ &= (1-\delta)[U(D)-X] \\ \frac{\partial R}{\partial U(C)} &> 0. \end{aligned}$$

Using Eqn. 7,

$$\frac{\partial R}{\partial U(C)} > 0 \Rightarrow \frac{\partial T}{\partial U(C)} < 0. \quad (10)$$

The number of punishment periods necessary to sustain cooperation varies inversely with the productivity of the union.

We can also find the relationship between the patience of the players and the number of punishment periods. Note since $R \geq \delta^{T+1}$, $\delta < R \Rightarrow T = 0$. Therefore, in all interesting cases, where $T > 0$, $R \leq \delta$. Coupled with the fact that $\delta < 1$, this implies

$$\ln\theta < \ln\delta < 0. \quad (11)$$

We are now ready to find $\frac{\partial T}{\partial \delta}$. Using Eqns 9 & 11, we can sign the derivative:

$$\begin{aligned} \frac{\partial T}{\partial \delta} &= \frac{(\ln \delta)(-\frac{1}{\delta}) - (\ln \delta)(\ln R - \ln \delta)}{[\ln \delta]^2} \\ \frac{\partial T}{\partial \delta} &< 0. \end{aligned} \tag{12}$$

The minimal number of punishment periods varies inversely with the discount factor.

Proof of Proposition 3: We restrict our attention to symmetric strategies. By using Eqn.s 4 and 5, we reconstruct V' to read:

$$V' = \frac{\frac{\partial U}{\partial a_i}\Big((1-\delta) + p(\delta - \delta^{T+1})\Big) - \frac{\partial p}{\partial a_i}\Big((U-X)(\delta - \delta^{T+1})\Big)}{((1-\delta) + p(\delta - \delta^{T+1}))^2} \tag{13}$$

If the probability of punishment and the utility from shirking are linear in the level of compliance, then we can write the derivatives of these functions as constants. Let $\frac{\partial U}{\partial a_i} = -\alpha$ and $\frac{\partial p}{\partial a_i} = -\beta$. We can now substitute these constants into Eqn. 13:

$$V' = \frac{(-\alpha)\Big((1-\delta) + p(\delta - \delta^{T+1})\Big) - (-\beta)(U-X)(\delta - \delta^{T+1})}{((1-\delta) + p(\delta - \delta^{T+1}))^2} \tag{14}$$

Recall that p and U are functions of a_i. We can show that

$$\frac{\partial V}{\partial a_i}|_{\vec{a}=1} > \frac{\partial V}{\partial a_i}|_{\vec{a}=0}$$

because $p(1) < p(0)$—the punishment regime is less likely to be triggered if all comply fully than if no one complies at all—and because by assumption, $U(1) > U(0)$. Therefore, we have three possible cases: (1) If $\frac{\partial V}{\partial a_i}|_{\vec{a}=1}$ is negative, then so is $\frac{\partial V}{\partial a_i}|_{\vec{a}=0}$. The derivative is always negative, for all $a_i \in [0,1]$. (2) If $\frac{\partial V}{\partial a_i}|_{\vec{a}=0}$ is positive, then so is $\frac{\partial V}{\partial a_i}|_{\vec{a}=1}$, and the derivative is always positive. (3) The derivative may start negative, $\frac{\partial V}{\partial a_i}|_{\vec{a}=0} < 0$, and end positive, at $\frac{\partial V}{\partial a_i}|_{\vec{a}=1} > 0$, and equal zero at some $a^* \in (0,1)$. In case (1), no participant ever wants to comply more; agents are limited by the boundaries on a_i, so the equilibrium is $a^* = 0$. Case (2) represents the full compliance

equilibrium; complying more is always better, so agents fully comply and $a^* = 1$. Case (3) allows for the possibility of an equilibrium with shirking. However, it can be shown[24] that the second derivative is zero: $\frac{\partial^2 V}{\partial a_i^2} = 0$, and we have an unstable interior equilibrium. Note that while it doesn't affect an individual's utility to increase compliance marginally, it does raise all other players' utility, because of the positive cross-partials: $\frac{\partial^2 V}{\partial a_i \partial a_j} > 0, \forall j \neq i$.

Therefore, the only stable equilibria are full compliance or no compliance.

Proof of Proposition 4: *Mas-Colell, Whinston, and Green, 1995, Proposition 8D3, p. 253.* A pure strategy Nash equilibrium exists in the game $[I, \{S_i\}, \{V_i(\cdot)\}]$, where I is the set of players, S_i a set of strategies for each player, and V_i is a utility function for each player, provided $\forall i$ (i) that the set of strategies S_i is a non-empty, convex, compact subset of some Euclidean space R^m, (ii) that $V_i(\cdot)$ is continuous in all players' strategies, and (iii) that $V_i(\cdot)$ is quasi-concave in s_i. Condition (iii) holds given Eqn. 6. The game described meets all of these conditions.

Proof of Proposition 5: Proposition 5 loosens the restrictions on the utility function from Proposition 3; we require only that the utility function be concave. Recall Eqn. 13:

$$V' = \frac{\frac{\partial U}{\partial a_i}\left((1-\delta) + p(\delta - \delta^{T+1})\right) - \frac{\partial p}{\partial a_i}\left((U-X)(\delta - \delta^{T+1})\right)}{((1-\delta) + p(\delta - \delta^{T+1}))^2}.$$

Proposition 5 is proven if (i) $\frac{\partial V}{\partial a_i} > 0$ at $\vec{a} = 0$ and (ii) $\frac{\partial V}{\partial a_i} < 0$ at $\vec{a} = 1$. Let $-\beta = \frac{\partial p}{\partial a_i}$ since the probability of punishment is linear. Substituting into Eqn. 13, the first condition is satisfied if:

$$\frac{\frac{\partial U}{\partial a_i}|_{\vec{a}=0}\left((1-\delta) + p(0)(\delta - \delta^{T+1})\right) + \beta\left((U(0)-X)(\delta - \delta^{T+1})\right)}{((1-\delta) + p(0)(\delta - \delta^{T+1}))^2} > 0$$

[24] In the linear case, the second derivative of the value function is written:

$$V'' = V' \frac{2p'(\delta - \delta^{T+1})}{(1+\delta) + p(\delta - \delta^{T+1})}.$$

Therefore, if $V' = 0, V'' = 0$.

at $\vec{a} = 0$. If $\frac{\partial U}{\partial a_i} \longrightarrow 0$ as $\vec{a} \longrightarrow 0$, then it can be shown that the inequality holds. At $\vec{a} = 1$, the necessary condition becomes:

$$\frac{\frac{\partial U}{\partial a_i}|_{\vec{a}=1}\left((1-\delta)+p(1)(\delta-\delta^{T+1})\right)+\beta\left((U(1)-X)(\delta-\delta^{T+1})\right)}{((1-\delta)+p(1)(\delta-\delta^{T+1}))^2}<0. \tag{15}$$

Since $p(1) > 0$ and is probably small, the inequality in Eqn. 15 holds if

$$\frac{\partial U}{\partial a_i}|_{\vec{a}=1} < -\beta\left(\frac{(U(1)-X)(\delta-\delta^{T+1})}{(1-\delta)}\right).$$

Therefore, at most $\frac{\partial U}{\partial a_i}|_{\vec{a}=1}$ need be less than (and restoring $\frac{\partial p}{\partial a_i}$): $\frac{\partial p}{\partial a_i}\left(\frac{\delta(U(1)-X)}{1-\delta}\right)$.

Proof of Proposition 6: Here we allow the probability of punishment to be non-linear, while we keep the utility of deviation linear. Letting $\frac{\partial U}{\partial a_i} = -\alpha$, Eqn. 13 becomes:

$$V' = \frac{-\alpha\left((1-\delta)+p(\delta-\delta^{T+1})\right)-\frac{\partial p}{\partial a}\left((U-X)(\delta-\delta^{T+1})\right)}{((1-\delta)+p(\delta-\delta^{T+1}))^2}. \tag{16}$$

As with Proposition 5, no full compliance equilibria exist if (i) $\frac{\partial V}{\partial a_i} > 0$ at $a_i = 0$ and (ii) $\frac{\partial V}{\partial a_i} < 0$ at $a_i = 1$. The second condition is satisfied if $\frac{\partial p}{\partial a_i}|_{\vec{a}=1} \longrightarrow 0$ as $\vec{a} \longrightarrow 1$. However, it may also be true that $\frac{\partial p}{\partial a_i}|_{\vec{a}=0} \simeq 0$, in other words, the probability might not be convex, but instead might follow an S-shape. In this case, a sufficient condition to sustain any compliance in equilibrium is if there exists some $\bar{a} \in (0,1)$ for which the slope is sufficiently steep. Formally (dropping the denominator from Eqn. 16),

$$\begin{aligned} 0 &< -\alpha\left((1-\delta)+p(\bar{a})(\delta-\delta^{T+1})\right)-\frac{\partial p}{\partial a}\left((U\bar{a}-X)(\delta-\delta^{T+1})\right) \\ \frac{\partial p}{\partial a_i}|_{\vec{a}=\bar{a}} &< -\frac{\alpha\left((1-\delta)+p(\bar{a})(\delta-\delta^{T+1})\right)}{(U(\bar{a})-X)(\delta-\delta^{T+1})} \end{aligned}$$

While $p(\bar{a}) < 1$, we can set it equal to 1 to make the r.h.s. as negative as possible. Then, it is sufficient for the derivative to be

$$\frac{\partial p}{\partial a_i} < \frac{\alpha(1-\delta^{T+1})}{(U(\bar{a})-X)(\delta-\delta^{T+1})},$$

where $\alpha = \frac{\partial U}{\partial a_i}$.

References

Abreu, Dilip, David Pearce, Ennio Stacchetti. 1986. "Optimal Cartel Equilibria with Imperfect Monitoring." *Journal of Economic Theory* 39:251-269.

Abreu, Dilip, David Pearce, Ennio Stacchetti. 1990. "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring." *Econometrica* 58(5): 1041-1063.

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

Axelrod, Robert and Ross A. Hammond. 2003. "The Evolution of Ethnocentric Behavior." Unpublished manuscript, Department of Political Science, University of Michigan.

Banks, Jeff, Charles Plott, and David Porter. 1988. "An Experimental Analysis of Unanimity in Public Goods Provision Mechanisms." *Review of Economic Studies* 55(182):301-322.

Bednar, Jenna. 2004. "Judicial Predictability and Federal Stability: Strategic Consequences of Institutional Imperfection." *Journal of Theoretical Politics* 16(4):423–446.

Bednar, Jenna. 1998. *The Federal Problem: The Political Economy of Federal Stability.* PhD Dissertation, Stanford University Department of Political Science.

Bendor, Jonathan and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action." *American Political Science Review* 81(1): 129-154.

Bendor, Jonathan and Piotr Swistak. 1997. "The Evolutionary Stability of Cooperation." *American Political Science Review* 91(2):290-307.

Bhaskar, V. and Eric van Damme. 2002. "Moral Hazard and Private Monitoring." *Journal of Economic Theory* 102(1):16-39.

Boyd, Robert. 1988. "Is the Repeated Prisoner's Dilemma a Good Model of Reciprocal Altruism?" *Ethology and Sociobiology* 9:211-222.

Brown Weiss, Edith and Harold K. Jacobson. 1998. *Engaging Countries: Strengthening Compliance with International Environmental Accords.* Cambridge, MA: MIT Press.

Calvert, Randall L. 1993. "Communication in Institutions: Efficiency in a Repeated Prisoner's Dilemma with Hidden Information." In W. Barnett, M. Hinich, and N. Schofield, eds., *Political Economy: Institutions, Information, Competition, and Representation.* Cambridge University Press.

Chayes, Abram and Antonia Handler Chayes. 1993. "On compliance." *International Organization* 47(2):175-205.

Dougherty, Keith L. 2001. *Collective Action under the Articles of Confederation.* Cambridge University Press.

de Figueiredo, Rui and Barry R. Weingast. 1997. "Self-Enforcing Federalism: Solving the Two Fundamental Dilemmas." *Journal of Law, Economics, and Organization* 21(1):103–35.

Downs, George W. and David M. Rocke. 1995. *Optimal Imperfection? Domestic Uncertainty and Institutions in International Relations.* Princeton, NJ: Princeton University Press.

Downs, George W, David M. Rocke and Peter N. Barsoom. 1996. "Is the Good News about Compliance Good News about Cooperation?" *International Organization* 50(3):379-406.

Fearon, James D. 1998. "Bargaining, Enforcement, and International Cooperation." *International Organization* 52(2):269-305.

Fearon, James D. and David D. Laitin. 1996. "Explaining Interethnic Cooperation." *The American Political Science Review* 90 (4):715-735.

Feinberg, Robert and Christopher Snyder. 2002. "Collusion with Secret Price Cuts: An Experimental Investigation." *Economics Bulletin* 3(6):1-11.

Fudenberg, Drew and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54(3):533-554.

Fudenberg, Drew, David Levine, and Eric Maskin. 1994. "The Folk Theorem with Imperfect Public Information." *Econometrica* 62(5):997-1039.

Gintis, Herbert. 2000. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction.* Princeton: Princeton University Press.

Green, Edward J. and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Information." *Econometrica* 52(1): 87-100.

Greif, Avner, Paul Milgrom and Barry R. Weingast. 1994. "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild." *The Journal of Political Economy* 102(4):745-776.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162:1243-1248.

Huberman, Bernardo A. and Natalie S. Glance. 1993. "Evolutionary Games and Computer Simulations." *Proceedings of the National Academy of Sciences* USA, 90:7716-7718.

Koremenos, Barbara, Charles Lipson and Duncan Snidal. 2001. "The Rational Design of International Institutions." *International Organization* 55(4):761-800.

Krasner, Stephen D. 1991. "Global Communications and National Power: Life on the Pareto Frontier." World Politics 43(3):336-66.

Kreps, David M. 1990. *A course in Microeconomic Theory.* Princeton, NJ: Princeton University Press.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research," in John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press, pp. 111-194.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory.* Cambridge: Oxford Univ Press.

Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. "The Role of Institutions in the Revival of Trade: The Medieval Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics* 2(March):1-23.

Morrow, James D. 1994. "Modeling the Forms of International Cooperation: Distribution versus Information." *International Organization* 48(3):387-423.

Morrow, James D. 2001. "The Institutional Features of the Prisoners of War Treaties." *International Organization* 55(4):971-992.

Niou, Emerson M.S. and Peter C. Ordeshook. 1998. "Alliances versus Federations: An Extension of Riker's Analysis of Federal Formation." *Constitutional Political Economy* 9:271-288.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge, MA: Harvard University Press.

Ostrom, Elinor. 1999. "Coping with Tragedies of the Commons." *Annual Review of Political Science* 2:493-535.

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

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review* 86(2):404-417.

Palfrey, Thomas R. and Jeffrey E. Prisbrey. 1997. "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review* 87(5):829-846.

Rosendorff, B. Peter and Helen Milner. 2001. "The Optimal Design of International Trade Institutions: Uncertainty and Escape." *International Organization* 55(4):829-858.

Saijo, Tatsuyoshi and Nakamura, Hideki. 1995. "The 'Spite' Dilemma in Voluntary Contributions Mechanism Experiments." *Journal of Conflict Resolution* 39(3):535-60.

Scholz, John T. and Mark Lubell. 1998. "Trust and Taxpaying: Testing the Heuristic Approach to Collective Action." *American Journal of Political Science* 42(2): 398-417.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *American Political Science Review* 94(4):819-835.

Weesie, Jeroen and Reinhard Wippler. 1987. "The Cumulation of Incompetence in Organizations." *Journal of Mathematical Sociology* 12:347-382.

Weingast, Barry R. 1997. "The Political Foundations of Democracy and the Rule of Law." *American Political Science Review* 91(2):245-263.