The Signaling Power of Sanctions in Social Dilemmas

Joël van der Weele*
Goethe University Frankfurt

Evidence from field and laboratory experiments indicates that a large fraction of the people behave like conditional cooperators in public good games. In this article, I investigate the implications of the existence of both conditional cooperators and egoists for optimal law enforcement strategies. When norms of cooperation exist between conditional cooperators, sanctions set by an authority can be lower than in a "Hobbesian" setting where everybody is egoistic. Moreover, if the authorities have private information about the fraction of egoists in society, low sanctions can be optimal because they signal that there are few defectors and thus "crowd in" trust and cooperation between agents. In social dilemmas where conditional cooperation is an important factor, as is the case in tax compliance, the model provides a rationale for the low observed sanctions in the real world and the mixed evidence on the effectiveness of deterrence. (*JEL* D83, J30, K42, M52)

> Laws are partly formed for the sake of good men, in order to instruct them how they may live on friendly terms with one another, and partly for the sake of those who refuse to be instructed, whose spirit can not be subdued, or softened, or hindered from plunging into evil.
>
> Plato—The Laws

1. Introduction

What determines cooperation in social dilemmas has been a core problem for social scientists since the beginning of the discipline. Ever since Hobbes in the 17th century threatened the infamous "war of all against all," the dominant strand of literature highlights the role of sanctions in coercing people to

*Department of Economics, Goethe University Frankfurt, Grüneburgplatz 1, D-60323 Frankfurt, Germany. Email: vdweele@econ.uni-frankfurt.de.

The author would like to thank Karl Schlag, Rick van der Ploeg, Sanne Zwart, Joel Sobel, Pascal Courty, Tobias Broer, Jan Potters, Bastiaan Overvest, many seminar participants, and some anonymous referees for useful comments. The main part of this research was conducted at the European University Institute in Florence on a grant from the Nuffic grant authorities.

The Journal of Law, Economics, & Organization, Vol. 28, No. 1,
doi:10.1093/jleo/ewp039
Advance Access publication December 30, 2009
© The Author 2009. Published by Oxford University Press on behalf of Yale University.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

cooperate. But contemporary empirical research shows that people manage to find ways to cooperate even without the presence of authorities. There is substantial evidence that society has a large proportion of so-called conditional cooperators: agents that condition the decision to cooperate on what they think others do. The existence of such agents means that collective action problems may be partly a matter of coordination, and substantial cooperation may be achieved without the need for much coercion. However, in the absence of high sanctions, a necessary condition for such cooperation is trust; the belief that others are willing to cooperate.

If society is indeed a heterogenous mix of egoists and conditional cooperators, a pressing and largely ignored question is how coercion and trust can be combined to induce cooperation. Specifically, one may ask if trust between agents is independent of the use of sanctions. This article offers an answer to this question by presenting a model in which trust and coercion interact in determining cooperation. It argues that there is a trade-off between sanctions and trust.

The point of departure of the model is a standard social dilemma or public good game. The game is played in a large population of heterogeneous agents: whereas some of them are selfish, others are conditional cooperators who do not mind contributing if sufficiently many others do so. Agents know their own type but not that of the other players. It can thus be rational to either cooperate or defect depending on a player's own type and the expectation of the type of the rest of the players. The model includes a government or principal, who is the only one who knows the distribution of agents' types in society and can alter the payoffs of the game by introducing sanctions for defection.

In this game, I show that in case the conditional cooperators coordinate on mutual cooperation, there is a unique class of perfect Bayesian equilibria in which the government sets high sanctions if there are many egoists in society and low sanctions if there are many conditional cooperators. This means that high sanctions give a negative signal and crowd out the belief that others are of a high type. Although this decreases the motivation of conditional cooperators to cooperate, there is no crowding out on the behavioral level because the coercive power of the sanctions compensates for the effect of decreased trust in others. However, the signaling effect of sanctions leads the government to set lower sanctions in equilibrium in order to crowd in trust between citizens.

The model has applications in social dilemmas in large-scale societies or organizations. An application to tax evasion is discussed in the last section. The model suggests that the reason why real-world policies of tax evasion often feature low sanctions is that governments rely on the reciprocal preferences of the taxpayers. It also provides a rationale for evidence that raising sanctions on tax evasion sometimes has very little or even a negative effect on tax evasion (Sheffrin and Triest 1992). Being tough on tax evasion sends a mixed message: although evaders are being punished, they must be numerous to be taken so seriously. Thus, the article emphasizes a balancing act that the government must perform: It must deter those who are, to speak with Plato, inclined to

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

"plunge into evil," while maintaining the good men's motivation to live on friendly terms.

2. Literature

There is an increasing amount of evidence for the existence of so-called conditional cooperators. A conditional cooperator is someone who will cooperate if she thinks others will do so as well. Fehr and Gächter (2000) and Gächter (2007) review the evidence on conditional cooperation from public good games and field experiments. They conclude that a large amount of studies find much more cooperation than standard economic theory allows for and that much of this cooperation is conditional on (expected) cooperation of others. However, there is substantial heterogeneity in these preferences for reciprocity or conditional cooperation. Fischbacher and Gächter (2009) among others provide evidence from laboratory experiments for the existence of a number of types whose behavior is stable across games. They find that close to 55% of their subjects act as conditional cooperators, 25% act as pure free riders, and the rest shows more complicated behavior that often resembles conditional cooperation in the relevant range of play. Another source of evidence for conditional cooperation comes from field experiments that study contribution levels to charities. The results of four studies surveyed in Gächter (2007) are that those subjects who received information that others contributed a lot also contribute a lot. For example, Frey and Meier (2004) find that students contribute significantly more to charity funds if they were told that others contributed more in the past.

The existence of conditional cooperators implies that trust is a crucial variable for cooperation. Without being overly sophisticated, we can define trust in a collective action setting as a person's *belief* that others in society are of a virtuous nature and will therefore cooperate (we provide a more detailed definition below). The literature on trust in economics has largely been concerned with the consequences of trust for the economy. However, the question of how beliefs are determined by institutional arrangements has received much less attention.

One strand of literature that does investigate the relation between beliefs and institutions are theories that combine the analysis of law and social norms (see for a survey, McAdams and Rasmusen 2007). These theories hold that official rules have an impact on behavior by changing people's expectation of what others do. Cooter (1998) argues that nondeterrent laws may create focal points and help people in this way to coordinate on efficient outcomes. Tyran and Feld (2006) show in an experimental setup that mild, nondeterrent laws can be effective in raising contributions in a public good game if they are the result of a public voting procedure. Such a procedure allows people to express their intentions to cooperate. However, Kahan (2005) argues that an informational effect of the introduction of laws can also be negative. Official incentives express information about the dominant social values and norms in society. Consequently, a blanket crackdown on defection by the

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

government in the form of high sanctions will give people the idea that noncooperation is the prevailing social norm. To the extent that people are conditional cooperators, this reduces their own willingness to cooperate. This dual role of incentives is the main message of this article. In our setup, incentives have the traditional motivational effect that economists take them to have but they also shape the perceptions of people about the conduct of others in society.

Direct evidence for this idea comes from Galbiati et al. (2009). They have used a laboratory experiment to test the idea that the introduction of sanctions may have negative effects on beliefs and therefore behavior. Their experiment is based on a minimum effort coordination game, a variant on the well-known stag-hunt game, in which both players have to select an "effort" level. Each player is rewarded proportionally to the minimum of the two efforts, although she has to pay a cost that is proportional to her own effort. Subjects were matched in groups of three, two of whom played the minimum effort game. The Nash equilibria in this game are pairs of identical effort levels, and higher effort pairs Pareto dominate lower effort pairs. Simultaneous to the effort choice, the experimenters elicited the beliefs of the players about the effort pair of the other player. After the players had played the game once, the third player in the group (the principal) was informed of both effort levels. He was the only one who received this information and was rewarded proportionally to the minimum effort in the group. In one treatment, this "principal" could then decide whether or not to introduce a small and costly sanction that penalized deviations from the maximum effort level. The first two players were then informed of this decision (but not of the first-round outcomes) and played the minimum effort game a second time. In another treatment, an identical sanction was implemented unconditionally by the experimenters.

The results show that when the sanction is introduced "exogenously" by the experimenters, the subjects react with more optimistic beliefs and higher effort levels. Sanctions help coordination. However, when the sanction is introduced by a third player with superior information, sanctions were less effective, especially among those who played high effort in the first round. The authors explain this by a "signaling effect": the optimistic player who played relatively high effort in the first round was alerted by the introduction of a sanction that the other player had not behaved well. Consequently, she was reluctant to scale up her effort herself.

This phenomenon falls into the more general category of "crowding out" of cooperative behavior by formal incentives. A number of experiments, both in the laboratory and in the field, document that sanctions for deviant behavior sometimes increase such behavior. In a field experiment, Gneezy and Rustichini (2000) consider 10 day-care centers in Haifa. In five of them, they introduce a fine for parents who pick up their children late. In these five centers, the number of latecomers went up significantly in the weeks after the introduction of the fines and stayed up relative to the control group even after the fines had been withdrawn.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

An increasing amount of studies document similar findings in social dilemma settings. Frey and Oberholzer-Gee (1997) find that people are *less* likely to accept siting of waste facilities in their neighborhood when they are offered substantial financial compensation for it. They use several indicators of "civic-mindedness" to predict individual choices whether to accept the facility. They find that when compensation is offered, civic mindedness is no longer a predictor of this choice. They conclude that the compensation reduces the feelings of civic duty of citizens, which is consistent with the idea in this article. Ostmann (1998) provides experimental laboratory results that show that external enforcement financed by experiment participants only reduces "harvests" in common pool problem by a small amount relative to a no-enforcement treatment. Frey and Jegen (2001) and Bowles (2008) present surveys of the rapidly expanding empirical literature in this field. There is also a growing literature on the potentially negative effect of fines on cooperative behavior in experimental principal-agent and labor market settings (see, e.g., Fehr and Schmidt 2007). The information transmission we are discussing in this article may extend to other technologies of social control. Cialdini (2003) shows that moral appeals to abstain from antisocial behavior fail if they mention explicitly that the behavior is common in the population.

Two theoretical articles present signaling models of crowding out that are related to the present article. They both do so in a principal-agent context. In Bénabou and Tirole (2003), the principal has more information about the characteristics of a job and the ability of an agent to do it than the agent himself. The incentives that the principal chooses to introduce are therefore a signal to the agent that he might not be suitable, which diminishes his motivation for the job.

The article closest to the present one is Sliwka (2007), who also considers a principal-agent context. In the model, there are three types of employees: altruists, who take into account the principal's payoff, egoists, who maximize their own material payoff, and conformists, who prefer to do whatever they think the majority does. Because preferences of conformists depend on their beliefs about others, this is a psychological game. In this setting, the introduction of tight control by the principal may signal to the conformists that most people are selfish and this in turn will cause them to lower their effort. The principal may thus choose to trust rather than control the agents.

Although the signaling effect in this article is similar to that in Sliwka (2007), the models differ substantially. Instead of focusing on the vertical principal-agent relation, I look at the effects of information transmission on the horizontal cooperation *between* agents in a public good game. In this context, the model is applied to a concrete technology of social control, namely official sanctions. My assumptions are more traditional than that in Sliwka (2007). First, I do not use a psychological game. Beliefs in my model do not induce a preference change but serve the more traditional role of anticipating payoffs. Finally, Sliwka (2007) assumes that there is a large proportion of unconditionally altruistic types in the population, an assumption that is rejected by the (experimental) evidence. I deviate from the standard homo economicus only by assuming the well-documented conditional cooperator.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

3. The Model

The model is a sequential game of costly signaling with three different kinds of players: agents, a principal, and nature. The principal can be a government or a manager, and the agents correspondingly citizens or employees. Applications exist in both public and organizational context, but throughout this article, I will frame the problem as a public one and use the words "government," "citizens," and "society."

The central idea is the following: The citizens play a public good game with incomplete information. In contrast to standard assumptions, some of the citizens are conditional cooperators, who contribute only if they think a sufficient number of others does so. Whether mutual cooperation can be an equilibrium thus depends on the distribution of the types of the players. The citizens do not know the distribution of types but have a common prior over the possible distributions.

Nature starts the game by determining the distribution of types (thus transforming the game into one of imperfect information). The government is the only player who observes this distribution. Its objective is to maximize contributions to the public good. To this end it chooses the level of sanctions for defection. The sanctions are observed by the citizens in the economy before they choose their own action. Since the government has more information than the citizens, the citizens may make inferences from the sanctions about the distribution of types in society. There is thus double-sided asymmetric information: citizens have private knowledge of their type and the government has private knowledge of the distribution of types. In Section 4, we derive the equilibria of the game and show that asymmetric information may lead the government to set lower sanctions in equilibrium.

3.1 Timing

The timing of the game is as follows:

1. Nature chooses the state of society characterized by the proportion of high types ω.
2. The government observes ω and decides on its policy g.
3. The citizens learn their own type and the government policy g, update their prior, and choose their contribution level $c \in \{0,1\}$.

3.2 Nature

At the beginning of the game, nature determines the types of all agents in society. With probability ω each agent is chosen to be a high type. Thus, ω is the proportion of conditional cooperators in society and $1-\omega$ is the proportion of egoists. This proportion is itself a random variable Ω of which nature determines the realization. The probability that nature picks a given ω is given by a uniform distribution with support on $[0,1]$. We call the distribution characterized by ω *the state of society.*

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

3.3 The Government

The government observes the state of society (but not the individual types of the citizens). Thus, ω is the "type" of the government. The motivation for this assumption is that governments or managers are in an advantageous position to collect information about their citizens or employees. Governments employ bureaucracies that collect statistics on the aggregate behavior of citizens and keep records of the amount of law violations. By making policy, they also gain information about the reaction of the citizens. Managers meet with employees in different departments of the firm and monitor productivity, working hours, and indices of their corporate culture. Although the assumption of perfect knowledge of the type distribution is obviously extreme, it is likely that the combination of these information sources lead governments to have superior knowledge about society than any individual would have.

The government's objective is to maximize cooperation by the citizens in the economy. The instrument to do so is the use of costly "sanctions" $g \geqslant 0$, a punishment on defection by the agents. (I will use the words "sanctions," "punishment," and "incentives" interchangeably.) The government's objective function is

$$W(m,g) = m - \alpha g. \tag{1}$$

Here, m is the proportion of contributors in society and $0 < \alpha < 1$ is a cost parameter. The motivation for the assumption that higher sanctions are more costly is that they involve higher practical expenditures necessary to administer punishment and raise the probability of detection (in the model, the government does not know the individual types of players but merely the proportion of high types).

One could argue that $\alpha = \alpha(m)$ since costs may depend on the number of people who are not cooperating. Then, since the numbers of defenders decreases with the level of deterrence, it is possible that costs *de*crease in the size of the sanction.[1] Treating α as a constant is clearly a simplification, but the assumption that $\alpha(m)$ is positive for all m is sufficient for the (qualitative) results in this article. This assumption is reasonable because there is a (moral) limit on the severity of punishment, reflecting the idea that in a liberal society the punishment should be proportional to the crime. Indeed, in the real world we do not observe the death penalty for stealing a loaf of bread, even if this would be the most efficient way to deter bread thieves. This means that raising deterrence will at least in part have to rely on raising the probability of detection, which is expensive. Furthermore, even if punishments could be raised to high levels, this does not necessarily mean that sanctioning is cheap. For example, although longer prison sentences may deter some offenders, they will involve higher expenditures for those that do get caught. Harsher sentencing policies also raise the cost of false convictions, an argument that has been used against the death penalty. Finally, to be credible, a government that imposes

1. I thank an anonymous referee for this suggestion.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

high sanctions would still need the capacity to carry out those sanctions for a large number of offenders or there may be equilibria in which many people defect and escape sanctioning.

Note that I do not necessarily interpret the sanctions as fines, and there are no revenues to the government from the sanctions. Although fines could be part of a sanctioning scheme, I want to focus purely on the deterring or Hobbesian effect of sanction and not on the revenue-raising aspect. Note also that sanctions (and their costs) are set before citizens choose their actions. This implicitly assumes commitment by the government to carry out the sanctions once they are in place. This is natural in a setting where sanctions are decided upon by politicians, and their execution and enforcement are subsequently carried out by the executive and judiciary branch of government.

Finally, the setup can easily be extended to include incentives in the form of subsidies or rewards. If the government has the possibility to reward cooperation with a costly subsidy, doing so would send the same signal as sanctioning defection: incentives are apparently necessary because there are many egoists. Any incentive scheme that is costly to the government and raises the citizens' expected utility of cooperation relative to that of defection sends such a signal.[2]

3.4 The Citizens

We assume that there is a countably infinite population of agents or citizens of measure 1, indexed $i = 1, 2, \ldots$. There are two types of citizens. A fraction ω is a so-called conditional cooperator or high type, the rests are egoists or low types. After having learned her type (but not ω) and the government policy g, each citizen chooses a contribution level $c \in \{0, 1\}$. The payoffs π^e of an egoistic citizen i are as follows:

$$\pi_i^e(c_i, m) = h(m) - c_i - g(c_i). \tag{2}$$

Here, $h(m)$ is the individual payoff from the public good, financed by the contributions. We assume that $h(m)$ is increasing in the proportion of contributors m. Because the population consists of an infinite number of agents, the individual contribution is so small relative to the population size that we disregard its impact on m. This approximation simplifies things substantially. The second term c_i is the individual contribution and $g(c_i)$ is the government sanction, which is imposed only if the agent defects:

$$g(c_i) = \begin{cases} 0 & \text{if } c_i = 1, \\ g & \text{if } c_i = 0. \end{cases}$$

2. Empirically, there is some conflicting evidence about whether fines have the same effect as rewards. Andreoni *et al.* (2003) find that punishments for low offers in a modified dictator game are somewhat more effective than rewards. A combination of punishments and rewards works even better. However, Fehr and Schmidt (2007) find that adding a fine to a reward scheme can crowd out cooperative behavior of an agent in a principal-agent relation.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

It is easy to see that equation (2) induces a social dilemma because in the absence of sanctions it is a dominant strategy for the egoists not to contribute. Egoists will only contribute if the sanctions that the government sets for noncontribution are high enough, that is, if $g \geqslant 1$.

The payoffs π^c of a conditional cooperator are given by

$$\pi_i^c(c_i, m) = \begin{cases} h(m) - c_i - g(c_i) & \text{if } m < \overline{m}, \\ h(m) - \theta c_i - g(c_i) & \text{if } m \geqslant \overline{m}. \end{cases} \tag{3}$$

Here, $\theta \in (0,1]$ and $0 < \overline{m} < 1$. If aggregate contribution levels are low, conditional cooperators have the same cost of contributing as egoists. However, if aggregate contribution levels are high, the cost of contributing for an conditional cooperator is lower than that of an egoist. In fact, egoists are a special case of conditional cooperators with $\theta = 1$. The type space can thus be written $\Theta = \{1, \theta\}$.

We can interpret the parameter θ as a "warm glow" from contributing that only arises when others contribute. The strength of this warm glow decreases in θ. When others do not contribute, the warm glow disappears because one rather feels like the only "sucker" who contributes. Such a conditional feeling of warm glow is also interpretable as a reciprocal preference. In any case, the particular specification of preferences is not intended as being especially realistic but rather as a simple or reduced form that generates conditional cooperation. As such, it is consistent with that of models that have more structural pretensions, such as the social preference models of Bolton and Ockenfels (2000) or Fehr and Schmidt (1999).

To see that these preferences generate conditional cooperation, we let $p = P(m \geqslant \overline{m})$ denote the subjective belief that at least the threshold fraction of people contributes and compute the expected utilities of contributing and defecting

$$E \max[\pi^c(1,m)] \geqslant E[\pi^c(0,m)],$$

$$p(h(m) - \theta) + (1-p)(h(m) - 1) \geqslant h(m) - g,$$

$$p \geqslant \frac{1-g}{1-\theta}. \tag{4}$$

In words, equation (4) says that in order for a conditional cooperator to contribute, the subjective belief that at least a fraction $\overline{m}$ will contribute will have to be high enough. The stronger the warm glow (the lower is θ) and the stronger the sanctions g, the lower such expectations need to be to induce contributions from the high types. Throughout the analysis, we apply the tiebreaking rule that an indifferent agent complies.

In sum, the game the agents are playing is a standard public good game with two twists. The first twist is that the government can introduce sanctions that punish defection. The second twist is that a fraction ω of the players have no dominant strategy. Instead, their best response depends on what they think other players will do.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

3.5 Trust

In this article, I claim that sanctions can crowd out trust. However, the definition of trust is a notorious source of conflict. In Section 1, I defined trust in passing as the *belief* that the other is a high type. A trusting act (in this case, contributing to the public good) is performed on the basis of this belief. One thinks the other will cooperate because her intentions or character are virtuous. Other definitions, like Hardin's notion of "'encapsulated interest" (Hardin, 1991), define trust more broadly as a situation where the trustor has reason to think that the trustee cooperates because her interests are aligned with her own. This definition includes situations where the trustee is expected to cooperate because of external enforcement. In this article, we stick with the first definition because we are interested in how people assess the likelihood that others cooperate when sanctions are low. That is, trust can exist only in a situation in which the trustor is at risk precisely because she does not know the character of the people she is facing. By contrast, we define as "confidence" the belief that the other will cooperate out of self-interest.

So defined, we can interpret trust as an "intrinsic motivation" for cooperation that can sustain cooperation when "extrinsic motivation," that is, sanctions, is low or absent. In the model, a certain amount of trust defined in this way is a necessary condition for a conditional cooperator to cooperate if $g < 1$. By "crowding out of trust," I mean that higher sanctions are associated with lower trust, that is, with a lower posterior probability of each agent that the other agents are of a high type.

4. Crowding Out of Trust

This section is structured as follows: We start by introducing some notation and terminology. To clear the way for the analysis of asymmetric information, we first derive equilibria in the simpler but instructive case of symmetric information. Proposition 3, the main result, characterizes the equilibrium under asymmetric information. All proofs are in the Appendix.

Let $g(\omega)$ denote the government policy and $s(\Theta, g)$ the strategy of a citizen of type Θ. Denote by $\mu(\omega|\Theta, g)$ the posterior probability distribution of a citizen of type Θ about the state of society ω and by $U(s, m, g, \Theta)$ the expected utility to a citizen of playing strategy s. We define an equilibrium as follows.

Definition 1. An equilibrium consists of a government strategy g: $[0,1] \to \mathbb{R}_+$, a strategy for each citizen s: $\Theta \times \mathbb{R}_+ \to \{0,1\}$, and a posterior belief of each agent about the true state of society μ: $[0,1] \times \Theta \times \mathbb{R}_+ \to [0,1]$ such that

$$g(\omega) \in \arg\max_{g \in \mathbb{R}} W(m, g),$$

$$s(\Theta, g) \in \arg\max_{s \in S} U(s, m, g, \Theta),$$

$\mu(\omega|\Theta, g)$ is updated by Bayes' rule whenever possible.

This definition corresponds to that of a perfect Bayesian equilibrium (pBe). We restrict the analysis to pure strategy equilibria and require that the

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

equilibrium satisfy the "intuitive criterion (IC)" of Cho and Kreps (1987), a standard refinement of Bayesian-Nash equilibrium.

4.1 Symmetric Information

Before we tackle the asymmetric information case, it will be instructive to discuss the case in which the citizens know ω. We solve the game backward, and start with the reaction function of the citizens. In the absence of high sanctions and if $\omega > \overline{m}$, conditional cooperators face a coordination game among themselves. There is an equilibrium in which they all contribute and one in which they all defect. The equilibrium of the larger game depends on the equilibrium in this coordination game. We will see that when high types coordinate on contribution, there is an unique equilibrium, which features two pooling regions. We develop some terminology for this partial pooling (or semi-separating) equilibrium. In this equilibrium, there are two regions of realizations of ω, in each of which the government plays the same policy. We call the equilibrium threshold value between the regions ω^*_{SI} (or ω^*_{AI} in the asymmetric information case). We call a region where $\omega \in [0, \omega^*_{\mathrm{SI}})$ (i.e., where society consists of relatively many egoists) a "bad state of society," and those where $\omega \in [\omega^*_{\mathrm{SI}}, 1]$ a "good state of society." We label the government policy for this partial pooling equilibrium as follows: the policy that is set in the bad state of society is called g_1, and the policy in the good state of society is called g_2.

Proposition 1. Under symmetric information, there are two pBe:

1. When high types coordinate on not contributing when $g < 1$, the unique equilibrium is a Hobbesian pooling equilibrium in which the government sets $g = 1$ and all citizens contribute. If the government were to set $g < 1$, all citizens would defect.
2. When high types coordinate on contributing, the unique equilibrium features a threshold ω^*. A government that observes $\omega < \omega^*_{\mathrm{SI}}$ sets a sanction $g^*_1 = 1$ and all citizens cooperate. A government that observes $\omega \geqslant \omega^*_{\mathrm{SI}}$ sets a sanction $g^*_2 = \theta < 1$ and only the high types cooperate.

As explained above, the conditional cooperators face a coordination game among themselves if $g < 1$. The conditional cooperators can coordinate either on mutual contribution or on mutual defection. We can interpret these equilibria as being associated with a social norm of contribution or a social norm of defection. The fraction of conditional cooperators determines the amount of norm adherence. The first part of Proposition 1 describes the Hobbesian pooling equilibrium in which high types coordinate on defection. In this case, high types are behaviorally equivalent to egoists, and it is perhaps unsurprising that the model generates a Hobbesian conclusion, which says that only strong punishment will induce agents to contribute.

The second part of the proposition tells us that when the high types coordinate on contribution, the government strategy has a threshold ω^*_{SI}. The intuition is again straightforward: government types below ω^*_{SI} will never set low

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

sanctions (<1) because there are too many egoists around. Inducing cooperation only from the high types generates so few contributions that it pays to set a high sanction. Government types above ω_{SI}^{*} will set low sanctions: because there are few egoists, low sanctions are a cheap way to induce a high level of contributions. Thus, when there are many conditional cooperators, and those conditional cooperators follow a norm of contributing, the government does best to implement low sanctions and tolerate a few defectors. Social norms are such that there is no reason for the government to use costly coercive strategies.

This simple setup captures two extremes in political thinking. On the one hand, when social norms of cooperation are absent, we are led to a Hobbesian conclusion. On the other hand, it shows that when there is a sufficient amount of people who follow a cooperative social norm, sanctions can be low. The latter is a simple consequence of the existence of conditional cooperators, and something we seem to observe in many real-world social dilemmas.

4.2 Asymmetric Information

We now turn to the case of asymmetric information in which the government is the only player who knows ω. To start with, we can immediately verify the existence of Hobbesian equilibrium, just as in the symmetric information case. The proof of the existence of this equilibrium did not depend on the information conditions. The reason is that when high types coordinate on defection, their beliefs about ω are irrelevant.

Proposition 2. Under asymmetric information about ω, there is a Hobbesian pooling pBe in which the government sets $g = 1$ and everyone contributes. If the government were to set $g < 1$, everyone would defect.

In the remainder of the article, we focus on equilibria in which high types coordinate on cooperation, that is, there is a norm for contribution. It turns out that under asymmetric information, the analysis is substantially more complicated if high types coordinate on cooperation. Before we characterize the equilibria of the game, we collect some useful results that serve to narrow down the search.

Lemma 1. In any pBe in which high types coordinate on cooperation, there are at most two different levels of sanctions g.

Lemma 1 narrows down the search substantially. It implies that there are only two possible types of equilibria in which high types coordinate on contributing: pooling equilibria and semi-separating (or partial pooling) equilibria with two pooling regions. Making use of the IC, the following lemma rules out the former.

Lemma 2. In a pBe in which high types coordinate on cooperation, there are no pooling equilibria that satisfy the IC.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

The intuition behind this lemma is the following: Governments that observe a very bad state of society will always set a high sanction. If they did not, the egoists who are a substantial part of the population would defect. On the other hand, governments that observe a very good state of society will always want to set a low sanction because this is a cheap way to induce cooperation of the great majority of people. This is not immediately obvious: one might think that there exist pooling equilibria on $g = 1$ supported by very pessimistic off-equilibrium beliefs. However, one can show that ruling out deviations to sanctions below $g = 1$ by governments that observed a very high ω requires off-equilibrium beliefs that are "unreasonable" (as judged by the IC). To rule out such deviations, off-equilibrium beliefs would have to be very pessimistic. However, a deviation to a low sanction is not attractive for governments who observe a high proportion of egoists since they would induce very little cooperation. Therefore, the only governments that can be reasonably expected to deviate are those who observe relatively many conditional cooperators. Thus, pooling equilibria based on such pessimistic beliefs do not survive the IC.

Summing up the results of our two lemmas, we know that an equilibrium should feature two pooling regions. We are now in a position to state the main result of this study.

Proposition 3. Under asymmetric information about ω,

1. if high types coordinate on cooperation, the unique class of pure strategy pBe that satisfies the IC has two pooling regions characterized by the parameter ω^*_{AI}. A government that observes $\omega < \omega^*_{\mathrm{AI}}$ sets a sanction $g^*_1 = 1$ and all citizens cooperate. A government that observes $\omega \geqslant \omega^*_{\mathrm{AI}}$ sets a sanction $g^*_2 < 1$ and only the high types cooperate.
2. If $\overline{m} \geqslant 1-\alpha(1-\theta)$, then under asymmetric information there exist equilibria in which the equilibrium threshold $\omega^*_{\mathrm{AI}} < \omega^*_{\mathrm{SI}}$.

The first part of Proposition 3 repeats the result of Proposition 1 that when there are many conditional cooperators, government does best to implement low sanctions and tolerate a few defectors. The intuition is straightforward: the government will punish heavily when it knows that there are a lot of egoists around because this is the only way to ensure substantial amounts of cooperation in such an environment. It will punish less heavily when it expects many citizens to follow a norm of conditional cooperation because cooperation can be induced cheaply in such an environment by setting lower sanctions. However, in contrast to the symmetric information case, such a government strategy implies crowding out of trust in equilibrium because higher sanctions transmit information about the state of society to the citizens. This means that sanctions are "bad news."

The second part of the proposition states the implication of this signaling effect for government policy. It says that there is a continuum of equilibria under asymmetric information in which the government plays low sanctions $g^*_2 < 1$ for values of ω where it would not do so under symmetric information.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

The intuition behind this result is that when there is a norm of contribution between the high types, the government induces trust of citizens by setting a low sanction. To see how this works, consider a government under symmetric information that observes a state of society $\omega < \overline{m}$. Under symmetric information, the citizens know that ω is the state of society and the high types will not be motivated to cooperate. However, under asymmetric information, agents are more optimistic in the sense that upon observing low equilibrium sanctions, they attach positive probability to states of society that are higher than $\overline{m}$. Because beliefs and sanctions are complements in generating compliance from the high types, this allows the government to set lower sanctions. In turn, lower sanctions make inducing cooperation cheaper, which expands the region in which the government plays low sanctions. Thus, low sanctions induce citizens to trust each other more and thereby they crowd in cooperation between the citizens.

Figure 1 shows the region in which the authorities play low sanctions under both symmetric and asymmetric information. In the gray area, low sanctions are played under symmetric information. The border of this area is the unique equilibrium threshold ω^*_{SI} for each level of $\overline{m}$. If we turn to asymmetric information, we see that if $\overline{m} < \omega$, that is, the threshold cooperation level to experience a warm glow is relatively low, the equilibria under symmetric and asymmetric information coincide. The reason is that when $\overline{m}$ is low, low sanctions do not make people more optimistic than they would be if they

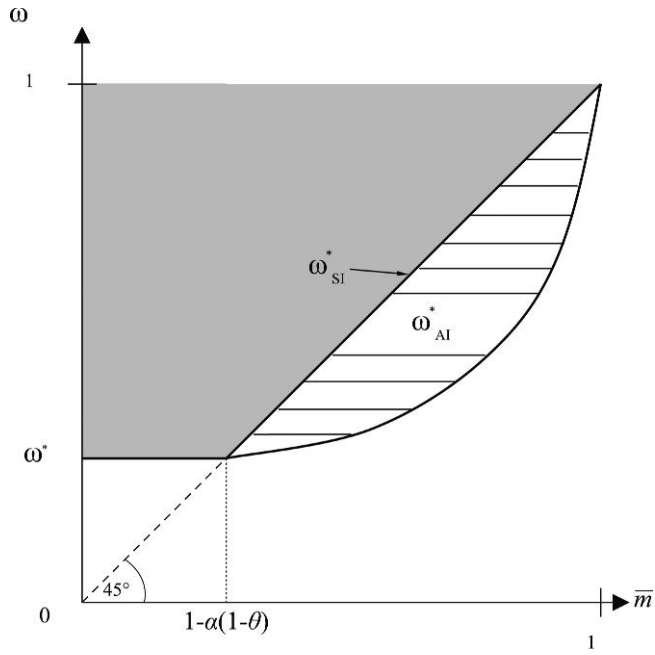

Figure 1. Equilibrium Region with Low Sanctions under (A)symmetric Information.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

knew ω. Note that equilibria with high levels of ω^*, supported by negative off-equilibrium beliefs, can be ruled out by the IC.

However, if $\overline{m} > \omega$, then the region where low sanctions are played expands under asymmetric information. In the hatched area, all values of ω^*_{AI} can be equilibrium values, and ω^*_{AI} can be lower than under symmetric information because of the "good news" effect explained above. The reason that there exist multiple equilibria when $\overline{m}$ is relatively high is that the lower bound of the reasonable (as judged by the IC) off-equilibrium beliefs is now lower than $\overline{m}$. This means that we can find off-equilibrium beliefs such that no one will cooperate when they see a deviation to a sanction lower than the equilibrium sanction. This supports the existence of many equilibria.

The comparative statics of α, the cost of sanctions, and θ, the strength of the warm glow, is intuitive. If the cost of sanctions increases, the region in which low sanctions are played increases. The same is true if the strength of the warm glow increases (θ decreases).

A final, rather subtle effect of asymmetric information is that high types are always more positive about the state of society than low types. An agent's own type gives her information about the state of society because the probability that each agent is a high type is given by ω. Thus, being a high type implies that others are more likely to be a high type.

In sum, asymmetric information enlarges the region where low sanctions can be played because low sanctions are good news. The signaling effect is a by-product of the fact that coercion is necessary only in bad states of society. In the terminology of Kahan (2005), it is truly the "expressive dimension" of sanctions.

5. Implications and Discussion

The model in this article can incorporate two extreme views of society. When $\theta = 1$ (no warm glow) and/or $\overline{m} = 1$, the agents in the model are all egoists and the model generates standard Hobbesian predictions. If $\overline{m}$ is low and $\theta = 0$, the model admits a rather romantic equilibrium in which equilibrium sanctions are zero: the government relies completely on social norms of cooperation. Realistically, the truth will be somewhere in between, so even if there are many conditional cooperators, the government still has a role to play. Although citizens' behavior is partly driven by trust, conditional cooperators will still need some motivation from an external sanctioning scheme because they are aware that there are some egoists around which reduces their desire to cooperate.

Second, there is no net crowding out of cooperation by sanctions. As in Bénabou and Tirole (2003), incentives are what they call "short-term reinforcers." In both models, higher sanctions "override" the effect of diminished beliefs. Thus, an econometrician looking solely at the relation between sanctions and cooperation would support the standard Becker-Stigler results. However, there is crowding out on the level of trust that influences the optimal sanction level. This brings us back to the definitions of "trust" and "confidence" as defined in Section 3. It should be clear that in contrast to trust,

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

confidence increases with sanctions because high sanctions make it in everybody's interest to cooperate.

Even though in this model sanctions compensate for the behavioral effects of decreased trust of agents, it suggests ways in which decreased trust may affect behavior. The model is consistent with the experimental observations of crowding out of cooperation: a drop in contributions if sanctions are raised. We must not forget that the sanctions implemented in (field) experiments are always off-equilibrium sanctions. They may thus interact with off-equilibrium beliefs. In this model, equilibria on low sanctions are supported by negative off-equilibrium beliefs. Thus, implementing a deviation to a (higher) off-equilibrium sanction may lead to less contributions.

Furthermore, trust is an attitude that determines behavior in many social situations. The crowding out of trust by incentives in one area could therefore have spillover effects in other policy areas and into the future. Suppose that besides playing the public good game described above, agents are matched privately with each other to play another dilemma or trust game. In each of those games, agents face partners drawn from the state of society. A government that sets a high sanction may improve cooperation levels in the public good game but will induce negative beliefs that may cause agents to defect in private interactions. Thus, a raise in sanctions in one policy area may cause a drop in cooperative behavior in other areas. As an example, consider the stigmatizing effect of police crackdowns on immigrant populations. This may lead people to think that immigrants must be criminal to have merited such police action. This may make them less willing to cooperate with immigrants in private interactions.

Sanctions may also have spillover effects into the future. Since the government cannot undo an information transmission, trust may not easily return. For example, when high sanctions are lowered after they have been introduced, for reasons not described in the model, cooperation may see a large drop, as even the by now cynical high types will refuse to cooperate. This is consistent with experimental evidence in, for example, Gneezy and Rustichini (2000): when incentives are withdrawn, cooperation does not return to preincentive levels.

A proper analysis of these ideas is a task for future research. Souvorov (2003) has worked in this direction and shows an intertemporal "addiction to rewards" in a two-period model of a principal and a single agent. In the context of our model, spillover effects will result in an "addiction to sanctions" as principals will need to maintain controlling measures to compensate for the reduced trust.

6. An Application to Tax Evasion

The potential applications of the model described in this article are everywhere where the conditions of the model are met: the principal has more information than the agents, some agents behave as conditional cooperators, and sanctions are costly. Kahan (2005) suggests applications in the public realm including not in my backyard problems and tax evasion (discussed below). One can also think of fare evasion in public transport, where the size of the penalty is an

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

indication of the norm of free riding. In the context of organizations and personnel economics, one can apply the model to incentive structures in large organizations and teams. In the context of sports, one can think of the doping dilemma, where harsh sanctions are indicative of a norm of widespread use of doping.

The example of tax evasion fits the model well because it is a private activity: any single taxpayer has very limited information on how honestly others pay their taxes. Tax offices on the other hand estimate evasion rates. This makes tax enforcement policies a vehicle of signals on how widespread tax evasion is. Moreover, there is overwhelming evidence that conditional cooperation is a prevalent attitude in tax compliance. Econometric studies conducted both on an individual level (Scholz 1998) and on an aggregate level (Frey and Torgler 2007) show that the decision to evade taxes is in large part based on dispositional attitudes. Especially important are the belief that fellow taxpayers evade and the perceived legitimacy of the use of tax revenue.

The model in this article can explain some puzzling facts about tax evasion. Andreoni et al. (1998: 821) remark that "For small amounts of evasion, [. . .] the expected cost of detection would appear to be extremely low for most taxpayers. So, we may ask, why are so many households honest, and why do not cheaters cheat by more?." The model in this article readily provides an answer to this question: people are conditionally cooperative, and as a consequence the government's best response is to apply mild (and cheap) sanctions instead of relying on heavy deterrence.

Another prediction of the model is that in equilibrium, low types pay taxes only for high sanctions, whereas high types will pay their taxes for a range of low sanctions. Wenzel (2004) shows in the context of tax evasion that official sanctions are effective only for those that have a weak personal norm of paying taxes. People with strong personal norms on the other hand also cooperate for low sanctions.

Evidence from (field) experiments also gives some indications that a signaling effect of sanctions is at work. Coleman (1997) reports the results of an experiment among 47,000 taxpayers in Minnesota. Some 1700 of them received a letter announcing that they had been randomly selected for an audit. The responses with respect to reported income were mixed: middle- and low-income taxpayers increased their reported income (although most of them by small amounts), but high-income taxpayers did not. In one treatment, the experimenters sent another letter to 20,000 taxpayers saying that the number of cheating taxpayers was much lower than commonly assumed. This significantly increased reported income. Sheffrin and Triest (1992) find that highly publicized campaigns against tax evasion often fail to have the desired effect and that some forms of information may increase distrust in other citizens.

7. Concluding Remarks

Polinsky and Shavel (2000), in their survey on theory of law enforcement, note that from a theoretical perspective sanctions often are too low. In their

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

conclusion, they remark that "Given the ample opportunities that exist for augmenting penalties, as well as the possible desirability of increasing enforcement effort, society should probably raise deterrence in many areas of enforcement." (2000: 72).

This article gives an explanation why sanctions may be "too low." It asks whether Hobbesian coercion in social dilemma problems remains optimal when society is a mix of conditional cooperators and egoists. What is the optimal policy to promote cooperation if the situation in question is a prisoners' dilemma for some and a coordination game for others? The article shows that the optimal level of sanctions depends on the relative proportions of the two agents in society. When there are many egoists, the high sanction or Hobbesian solution is optimal. When there are many conditional cooperators, a policy of low sanctions may be more efficient. If the government knows more about the composition of types in society, this implies that high sanctions are bad news. Thus, its superior information allows government to induce or crowd in cooperation by setting low sanctions. The article thus shows that sanctions may have a dual role. They both change economic payoffs and alter agents' perception of the environment. The government has to perform a balancing act: it has to punish the deviators, while keeping the conditional cooperators optimistic.

Appendix

Proof of Proposition 1.

1. We work backward through the game and start by characterizing the agents' reaction functions. We know from equations (2) and (4) that both types have a "threshold sanction": for lower sanctions than this threshold they defect, for higher sanctions they cooperate. Low types cooperate when the sanction is higher than 1 and defect otherwise. From equation (4), we know that high types cooperate when $g \geqslant 1-(1-\theta)p(m > \overline{m})$ and defect otherwise. In the symmetric information when $\geqslant \overline{m}$, it is sufficient that

$$g \geqslant \theta. \tag{A.1}$$

The reaction functions imply that when $g < 1$, all egoists defect and the conditional cooperators face a coordination game between themselves. Suppose high types coordinate on defection. In this case, the government can set $g < 1$ resulting in $m = 0$ or it can set $g = 1$ resulting in $m = 1$. From the objective function of the government it is straightforward to verify that when $\alpha < 1$, the latter strategy dominates the former.

2. Above we derived the reaction functions of the citizens. We know that equation (A.1) holds with equality in equilibrium, so that $g_2^* = \theta$ because the government always sets the lowest possible sanctions to induce cooperation. The government will set low sanctions iff

$$\begin{aligned} W(\omega, g_2^*) &\geqslant W(1, g_1^*), \\ \omega - \alpha g_2^* &\geqslant 1-\alpha, \\ \omega &\geqslant 1-\alpha(1-g_2^*). \end{aligned} \tag{A.2}$$

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

In equilibrium, this "incentive compatibility constraint" holds with equality for the lowest government type that sets low sanctions and with inequality for all higher types. Since the government will always set the lowest possible sanctions in equilibrium, that is, $g_2^* = \theta$, the threshold government type is given by $1-\alpha(1-\theta)$.

Naturally, government will set high sanctions if $\omega < \overline{m}$. Thus, we have

$$\omega^* = \begin{cases} 1-\alpha(1-\theta) & \text{if } \overline{m} < 1-\alpha(1-\theta), \\ \overline{m} & \text{if } \overline{m} \geqslant 1-\alpha(1-\theta). \end{cases} \tag{A.3}$$

□

Proof of Proposition 2. Identical to that of Proposition 1, Part 1. □

Proof of Lemma 1. From the reaction functions derived above, we see that contribution by the low types implies contribution by the high types. Thus, there are at most three different equilibrium action profiles for the citizens in the economy: one where both types contribute, one where only the high types contribute, and one where nobody contributes. This means that in equilibrium there are at most three different levels of sanctions g. If there were more, two such levels induce the same strategic reactions by the agents. This cannot be an equilibrium since the government would always deviate to the lower and cheaper sanction that induces a given reaction. Moreover, since $\alpha < 1$, we see from the welfare function that setting $g = 1$ and inducing full cooperation always yield a higher payoff to the government than setting $g = 0$ and leaving everybody to defect. Thus, defection by all cannot be an equilibrium outcome. We are left with at most two possible equilibrium outcomes: one where both types contribute, one where only the high types contribute. As a consequence, there are at most two sanction levels, one associated with each outcome. □

Proof of Lemma 2. We prove the lemma by showing the following:

1. A government that observes $\omega = 0$ sets $g = 1$ in equilibrium. This rules out any pooling equilibrium on $g < 1$.
2. For a government that observes $\omega = 1$, the upper bound on the equilibrium sanction is $\max\left\{\theta, 1-\frac{1}{\alpha}(1-\overline{m})\right\} < 1$. This rules out any pooling equilibrium on $g = 1$.

Proof of 1. In a state of society $w = 0$ where everybody is egoistic, setting $g < 1$ will lead everyone to defect that cannot be optimal for the government.

Proof of 2. The proof is based on the application of the IC (Cho and Kreps 1987), a refinement of Bayesian-Nash equilibrium. An equilibrium fails the IC if it requires off-equilibrium beliefs that place positive probability on types for whom deviation payoffs are dominated by equilibrium payoffs. The idea is that it is "unreasonable" to believe that such types would have deviated. Denote by $\Omega(g')$, the set of government types who will deviate to an off-equilibrium

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

sanction g'. We call beliefs with full density inside $\Omega(g')$ "IC-admissible." Then $[0,1]/\Omega(g')$ is the set of types who would never deviate to a sanction g'. Beliefs with density in this set are "non–IC-admissible."

We make two observations that restrict the set of deviations that we need to consider. First, we already ruled out pooling equilibria on $g < 1$, so we consider only deviations from $g = 1$. For a deviation to a sanction level $g < 1$, the contribution level will be either ω or 0. A deviation cannot be profitable if contributions are 0. Thus, we focus on deviations to sanctions g' that induce a contribution level of ω. From equation (A.1), we know that if $g' < \theta$, sanctions will never (for any beliefs) induce cooperation from high types, and so we look only at deviations to sanctions $\theta \leqslant g' < 1$.

Second, we can restrict our attention to deviations by the government type $\omega = 1$. In this case, the whole population consists of high types, and a contribution level of ω equals the maximum contribution level. Therefore, if this type does not deviate, other types will not do so either.

In sum, a pooling equilibrium on $g = 1$ exists if and only if for $\omega = 1$ and for all $\theta \leqslant g' < 1$, there exist off-equilibrium beliefs that are (a) IC-admissible, and (b) lead to zero contributions, thus making deviations unprofitable.

The set $\Omega(g')$ of government types that will deviate under a deviation is determined by comparing the government's utility in equilibrium to that of a deviation:

$$EW(\omega, g') \geqslant EW(\omega, g = 1),$$

$$\omega - \alpha g' \geqslant 1 - \alpha,$$

$$\omega \geqslant 1 - \alpha(1 - g').$$

Thus, we have $\Omega(g') = [1 - \alpha(1 - g'), 1]$. The best case for a pooling equilibrium is made when off-equilibrium beliefs are as low as possible given the IC, that is, have full density on $1 - \alpha(1 - g')$. These beliefs will lead to zero contributions if $\overline{m} > 1 - \alpha(1 - g')$. Solving for g' yields

$$g' < 1 - \frac{1}{\alpha}(1 - \overline{m}). \tag{A.4}$$

Thus, if an off-equilibrium sanction satisfies $g' \geqslant 1 - \frac{1}{\alpha}(1 - \overline{m})$, then there are no off-equilibrium beliefs that are IC-admissible and lower than $\overline{m}$.

Since $\overline{m} < 1$ and $\theta < 1$, we can always find sanctions $\max\{\theta, 1 - \frac{1}{\alpha}(1 - \overline{m})\} \leqslant g' < 1$, which means that a deviation to g' leads all citizens to cooperate. Thus, for the type $\omega = 1$, there is a profitable deviation to a sanction that is slightly lower than 1 and a pooling equilibrium on $g = 1$ cannot exist. □

Proof of Proposition 3. Note that for various reasons, we cannot use a standard single crossing property condition. The proof proceeds in four steps. First, we characterize the citizens' posterior belief about the distribution of types in the economy. Agents base their beliefs on the government's policy and their own type. We derive only the posterior beliefs of conditional cooperators

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

(high types) under a sanction $g < 1$ because this is the only case in which beliefs matter for the choice of action.[3]

Conditional on $g_2 < 1$ and $\Theta = \theta$, we compute from Bayes' rule the posterior belief distribution $\mu(\omega)$ that a given distribution ω has been chosen by nature. The common prior is that each distribution is equally likely to be chosen by nature. Obviously $\mu(\Omega = \omega < \omega^* | g = g_2, \Theta = \theta) = 0$ because the agent knows that a low sanction is played only if $\omega \geqslant \omega^*$. The posterior for $\Omega = \omega \geqslant \omega^*$ is

$$\begin{aligned}\mu(\Omega = \omega \geqslant \omega^* | \Theta = \theta, g = g_2) &= \frac{P(\Omega = \omega \cup \Theta = \theta \cup \omega \geqslant \omega^*)}{P(\Theta = \theta \cup \omega \geqslant \omega^*)} \\ &= \frac{\frac{\omega}{1-\omega^*}}{\int_{\omega^*}^{1} \frac{\omega}{1-\omega^*} d\omega} \\ &= \frac{2\omega}{1-(\omega^*)^2}.\end{aligned}$$

Hence,

$$\mu(\Omega = \omega | \Theta = \theta, g = g_2) = \begin{cases} 0 & \text{if } \omega < \omega^*, \\ \frac{2\omega}{1-(\omega^*)^2} & \text{if } \omega \geqslant \omega^*. \end{cases} \tag{A.5}$$

Second, we determine the best response of the citizens in the economy to any government policy given their posterior beliefs and their type. Both types will cooperate under $g_1 = 1$. We know that the best response of a low type is to defect whenever $g < 1$. Remains to analyze the case of a high type who observes g_2. From equation (4), we know that best response of a high type is to cooperate if and only if $P(m > \overline{m}) \geqslant \frac{1-g}{1-\theta}$.

To get the best response of the citizens, we have to compute the equilibrium value $P^*(m > \overline{m} | g_2^*)$ from the equilibrium beliefs. If $\overline{m} \leqslant \omega^*$, it is straightforward that $P^*(m > \overline{m} | g_2^*) = 1$. Substituting this in equation (4) yields the equilibrium condition for the cooperation of high types

$$g_2^* \geqslant \theta. \tag{A.6}$$

If $\overline{m} > \omega^*$, the equilibrium beliefs are given by the following equation:

$$\begin{aligned}P^*(m > \overline{m}) &= \int_{\overline{m}}^{1} \frac{2\omega}{1-(\omega^*)^2} d\omega \\ &= \frac{1-\overline{m}^2}{1-(\omega^*)^2}.\end{aligned}$$

Substituting this in equation (4) yields the equilibrium condition for the cooperation of high types:

$$g_2^* \geqslant \frac{\overline{m}^2 - (\omega^*)^2 + \theta(1-\overline{m}^2)}{1-(\omega^*)^2}. \tag{A.7}$$

3. Concerns of space lead me to omit the full characterization of posterior beliefs of agents. These are available on request.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

Third, the best response of the government types is described by the incentive compatibility constraint (A.2) derived above that gives the threshold type that is indifferent between the high and the low sanction.

The fourth step is deriving the equilibrium conditions on the parameter values starting with the equilibrium sanction. We need to consider both the case when $\overline{m} > \omega$ and the complement.

Case 1: $\overline{m} \leqslant \omega^*$.

In this case, equilibrium beliefs $P(m > \overline{m}) = 1$, and so from equation (4), it follows that $g_2^* \geqslant \theta$ is sufficient for cooperation of the high types. From the incentive compatibility constraint of the government (equation A.2), it follows that $\omega^* \geqslant 1-\alpha(1-\theta)$. It is easy to check that $g_2^* = \theta$ and $\omega^* = 1-\alpha(1-\theta)$ is an equilibrium as long as $\overline{m} \leqslant 1-\alpha(1-\theta)$. Deviations to $g_2 > \theta$ are never profitable and deviations to $g_2 < \theta$ lead to $m = 0$.

Now suppose that $g_2^* > \theta$. Consider a deviation to $g' = \theta$. The intuitive criterion specifies (see proof of Lemma 2) that the lowest reasonable off-equilibrium beliefs are $1-\alpha(1-g')$. A profitable deviation to $\theta \leqslant g' < g_2^*$ can be ruled out only if $\overline{m} > 1-\alpha(1-\theta)$. In this case, the equilibrium must satisfy $\omega^* = \overline{m}$ since otherwise there exists a profitable deviation to $g' < g_2^*$ such that $\omega^* = 1-\alpha(1-g_2^*) > 1-\alpha(1-g') > \overline{m}$.

Case 2: $\overline{m} > \omega^*$.

In this case, $P(m > \overline{m}) = \frac{1-\overline{m}^2}{1-(\omega^*)^2}$ and $g_2^* \geqslant \frac{\overline{m}^2-(\omega^*)^2+\theta(1-\overline{m}^2)}{1-(\omega^*)^2}$. From the government's incentive constraint (A.2) one can derive that the (lower bound of the) equilibrium threshold $\underline{\omega}$ is given implicitly by

$$(1-\underline{\omega})(1-\underline{\omega}^2) \leqslant \alpha(1-\theta)(1-\overline{m}^2). \tag{A.8}$$

Suppose that $g_2^* = \frac{\overline{m}^2-(\omega^*)^2+\theta(1-\overline{m}^2)}{1-(\omega^*)^2}$, so $\omega^* = \underline{\omega}$ and is given by equation (A.8) with equality. It is clear that deviations to $g' > g_2^*$ are never profitable. Deviations to $\theta \leqslant g' < g_2^*$ are unprofitable as long as $\overline{m} > 1-\alpha(1-g')$ (see proof of Lemma 2). We know from equation (A.2) that $\underline{\omega} > 1-\alpha(1-g')$. Thus, we have that $\overline{m} > \underline{\omega} > 1-\alpha(1-g')$. This means we can always find off-equilibrium beliefs that make a deviation unprofitable and the equilibrium exists.

Now consider as an equilibrium sanction $g_2^* > \frac{\overline{m}^2-(\omega^*)^2+\theta(1-\overline{m}^2)}{1-(\omega^*)^2}$, and thus (by equation (A.2)), $\omega^* > \underline{\omega}$. It is clear that deviations to $g' > g_2^*$ are never profitable. Deviations to $\theta \leqslant g' < g_2^*$ can be ruled out by reasonable off-equilibrium beliefs by the same reasoning as above. Thus, this equilibrium exists.

Summarizing, we have

$$\omega^* \begin{cases} = 1-\alpha(1-\theta) & \text{if } \overline{m} < 1-\alpha(1-\theta), \\ \in [\underline{\omega}, \overline{m}] & \text{if } \overline{m} \geqslant 1-\alpha(1-\theta), \end{cases}$$

where $\underline{\omega}$ is given in equation (A.8).

Proof of 2. Comparing ω^* under asymmetric information, with ω^* under symmetric information, the proof is immediate. □

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

References

Andreoni, J., B. Errard, and J. Feinstein. 1998. "Tax Compliance," 36 *Journal of Economic Literature* 818–60.

Andreoni, J., W. Harbaugh, and L. Vesterlund. 2003. "The Carrot or the Stick: Rewards, Punishments and Cooperation," 93 *American Economic Review* 893–902.

Bénabou, R., and J. Tirole. 2003. "Intrinsic and Extrinsic Motivation," 70 *Review of Economic Studies* 489–520.

Bolton, G. E., and A. Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity and Competition," 90 *American Economic Review* 166–93.

Bowles, S. 2008. "Policies Designed for Self-interested Citizens May Undermine the Moral Sentiments: Evidence from Economic Experiments," 320 *Science* 1605–09.

Cho, I.-K., and D. M. Kreps. 1987. "Signaling Games and Stable Equilibria," 102 *Quarterly Journal of Economics* 179–221.

Cialdini, R. B. 2003. "Crafting Normative Messages to Protect the Environment," 12 *Current Directions in Psychological Science* 105109.

Coleman, S. 1997. "Income Tax Compliance: A Unique Experiment in Minnesota," 13 *Government Finance Review* 11–5.

Cooter, R. 1998. "Expressive Law and Economics," 27 *Journal of Legal Studies* 585–608.

Fehr, E., and K. Schmidt. 1999. "A Theory of Fairness, Competition and Cooperation," 114 *Quarterly Journal of Economics* 817–68.

———. 2007. "Adding a Stick to the Carrot. The Interaction of Bonuses and Fines," 97 *American Economic Review, Papers and Proceedings* 177–81.

Fehr, E., and S. Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity," 14 *Journal of Economic Perspectives* 159–81.

Fischbacher, U., and S. Gächter. 2009. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods," *American Economic Review* (forthcoming).

Frey, Bruno S. and Reto Jegen. 2001. "Motivation Crowding Theory," 15 *Journal of Economic Surveys* 589–611.

Frey, B. S., and B. Torgler. 2007. "Tax Morale and Conditional Cooperation," 35 *Journal of Comparative Economics* 136–59.

Frey, B., S. and F. Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-out," 87 *American Economic Review* 746–55.

Frey, B. S., and S. Meier. 2004. "Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment," 94 *American Economic Review* 1717–22.

Gächter, S. 2007. "Conditional Cooperation. Behavioral Regularities from the Lab and the Field and their Policy Implications," in Bruno S. Frey and Alois Stutzer, eds., *Economics and Psychology. A Promising New Cross-Disciplinary Field.* CESifo Seminar Series. Cambridge, MA: MIT Press.

Galbiati, R., K. H. Schlag, and J. van der Weele. 2009. "Can Sanctions Induce Pessimism? An Experiment," Labsi Working Paper 24/2009. University of Siena.

Gneezy, U., and A. Rustichini. 2000. "A Fine is a Price," 29 *Journal of Legal Studies* 1–17.

Hardin, R. 1991. "Trusting Persons, Trusting Institutions," in Richard J. Zeckhauser, ed., *The Strategy of Choice*. Cambridge, MA: MIT Press.

Kahan, D. M. 2005. "The Logic of Reciprocity: Trust, Collective Action, and Law," in Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr, eds., *Moral Sentiments and Material Interests.* Cambridge, MA: MIT Press.

McAdams, R. H., and Rasmusen, E. B. 2007. "Norms in Law and Economics," in A. M. Polinsky and S. Shavell, eds., *The Handbook of Law and Economics*. North Holland, pp. 1573–628.

Ostmann, A. 1998. "External Control May Destroy the Commons," 10 *Rationality and Society* 103–22.

Polinsky, M. A., and S. Shavell. 2000. "The Economic Theory of Public Enforcement of Law," 38 *Journal of Economic Literature* 45–76.

Scholz, J. T. 1998. "Trust, Taxes, and Compliance," in Valerie Braithwaite and Margaret Levi, eds., *Trust and Governance*. New York, NY: Russell Sage Foundation.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013

Sheffrin, S. M., and R. K. Triest. 1992. "Can Brute Deterrence Backfire? Perceptions and Attitudes in Taxpayer Compliance", in Joel Slemrod, ed., *Why People Pay Taxes*. Ann Arbor, MI: University of Michigan Press.

Sliwka, D. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," 97 *American Economic Review* 999–1012.

Souvorov, A. 2003. "Addiction to Rewards," Toulouse, France: Mimeo GREMAQ.

Tyran, J.-R., and L. P. Feld. 2006. "Achieving Compliance when Legal Sanctions are Non-deterrent," 108 *Scandinavian Journal of Economics* 135–56.

Wenzel, M. 2004. "The Social Side of Sanctions: Personal and Social Norms as Moderators of Deterrence," 28 *Law and Human Behavior* 547–67.

Downloaded from http://jleo.oxfordjournals.org/ at Universiteit van Amsterdam on December 5, 2013