

# The Signalling Power of Sanctions in Collective Action Problems\*

Joël van der Weele  
European University Institute

June 1, 2007

## Abstract

We present a model of collective action in a heterogenous population of egoists and conditional cooperators. Each player is uncertain about the cooperative inclinations of the other player. A government or principal who has information about the distribution of types may introduce sanctions for defection. We study the impact of such sanctions through the effect on the beliefs of the players about the distribution of types they are facing. It is shown that in equilibrium sanctions can crowd out trust between agents by sending a signal that there are many egoists around. This can lead the government to set low sanctions to induce trust and ‘crowd in’ cooperation. In cases where conditional cooperation is an important factor in collective action, as is the case in tax compliance, the model provides a rationale for the low observed sanctions in the real world.

**Keywords:** Collective action, trust, incentives, crowding out, conditional cooperation

**JEL codes:** 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 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 collective action problems has been a core problem for social scientists since the beginning of the discipline. Ever since Hobbes threatened the infamous ‘war of all against all’ in the 17th century, the dominant strand of literature highlights the role of sanctions in coercing people to cooperate. But contemporary empirical research shows that people manage to find ways to cooperate even without the presence of government. 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. Collective action problems between conditional cooperators are therefore a matter of coordination rather than coercion. This puts trust in the centre of attention for research into collective action problems.

Thus, 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. How do sanctions and trust

---

\*I would like to thank Rick van der Ploeg for his support, Sanne Zwart, Joel Sobel, Pascal Courty, Tobias Broer, Javier Rivas, Cor van der Weele, Ken Binmore, Bastiaan Overvest, Mark LeQuement, and Anindya Banerjee for useful comments, and above all Karl Schlag for his commitment.

relate to each other, and how should the government optimally use sanctions when society consists of both egoists and conditional cooperators?

This paper offers an answer to these questions 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. High sanctions are necessary when there are many egoist around, but they may also ‘crowd out’ trust, because conditional cooperators will infer that being selfish is the norm. This in turn decreases the willingness of conditional cooperators to cooperate. As a consequence, the government can ‘crowd in’ cooperation by setting low sanctions. Because it has superior information about the distribution of types in society, the government can use the low sanction policy to signal a social norm of cooperation.

The point of departure of the model is a population of heterogeneous agents: While some of them are selfish, others are conditional cooperators. The agents are randomly matched in a 2 person game that, depending on the cooperative preferences of the participants, is either a prisoners dilemma or a coordination game. They know their own type, but not that of the other player. 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 other player. We include a government, that knows the distribution of agent’s types in society, and that can alter the payoffs of the game by introducing sanctions for defection.

We show that if conditional cooperators can coordinate on mutual cooperation, there is a unique perfect Bayesian equilibrium in which the government sets high sanctions when there are many egoists in society, and low sanctions when there are many conditional cooperators. This means high sanctions give a negative signal to the conditional cooperators and crowd out trust between citizens. Furthermore, the asymmetric information about distribution of types in society implies that average sanctions in society are lower than what they would be if citizens knew the distribution. The reason is that the government can induce trust and thereby ‘crowd in’ cooperation from it citizens by setting low sanctions. We also characterize the welfare costs incurred by a ‘Hobbesian’ government, that acts upon the idea that all people are egoists. Such a policy is unnecessarily costly when there are many conditional cooperators.

Benabou and Tirole (2003) present a model of crowding out in the context of a principal and a single agent. Sliwka (2007) considers a signalling effect of incentives in a principal-agent context. He models a psychological game where agents change their preferences depending on the policy they observe. We build on the approach of Benabou and Tirole, but focus on a collective action problem. That is, on the effect of the signal on the behavior of agents towards each other, rather than towards a principal. Furthermore, in contrast to Sliwka, we do not use a psychological game. Instead we achieve the result by adding heterogeneity of types to a standard collective action framework.

The model has applications in collective action problem in large scale societies or organizations. Perhaps the clearest application is tax evasion. Being tough on tax evasion sends a mixed message: although evaders are being punished, they must be numerous to be taken so seriously. The model asserts that the reason why real-world policies of tax evasion often feature low sanctions, is that governments rely on the reciprocal preferences of the tax-payers. The model suggest a rationale for evidence that raising sanctions on tax evasion has very little, or even a negative effect on tax evasion (Sheffrin and Triest, 1992). Other applications are support for the welfare state, free-riding in public transportation or team work in organizations. In all these cases, the article tries to shed light on a balancing act that the principal must perform: It must deter those who are, to speak with Plato, inclined to ‘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 (2006) review the evidence on conditional cooperation from public good games and field experiments. They conclude that a large amount of studies finds much more cooperation than standard economic theory allows for, and that this evidence can be comfortably explained by reciprocal preferences. However, there is substantial heterogeneity in these preferences for reciprocity or conditional cooperation. Fischbacher and Gächter (2006) among others, provide experimental evidence for the existence of a number of stable types. 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 (2006) are that those subjects who received information that others contributed a lot also contribute a lot. For example, Frey and Meyer (2004) find that students contribute significantly more to charity funds and have higher expectations of future contribution levels by others, if they were told that others contributed more in the past.

The existence of conditional cooperators implies that trust is 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 trustworthy. 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 idea that has gained ground is that sanctions have an impact on trust. Kahan (2005) calls the idea that incentives have social meaning the *expressive dimension* of incentives. Incentives express information about the social values and norms in society. Consequently, Kahan argues, a blanket crackdown on defection by the government in the form of high sanctions will give people the idea that non-cooperation 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 paper. 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.

This phenomenon falls into the category of what is commonly called 'crowding out of intrinsic motivation'. The debate about the adverse effects of incentives in psychology has been going on for decades, and in the last ten years it has spilled over to economics. The point of the debate is that material incentives appear to have more complex effects than just their impact on a person's economic trade-offs. If people have an 'intrinsic' motivation to perform a certain task, this motivation can be enhanced (crowded in) or diminished (crowded out) by external incentives. This effect has been dubbed 'motivation crowding', and can lead a person to react differently to incentives than standard economic theory predicts. Many studies have found these crowding effects in economic settings (see Frey and Jegen (2001) for an overview).

A beautiful illustration is provided by the widely cited paper by Gneezy and Rustichini (2000a). In their field experiment they consider ten 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 late-comers went up significantly in the weeks after the introduction of the fines and stayed up relative to the control group after the fines had been withdrawn.

An increasing amount of studies documents similar findings in collective action problems. A field experiment among Colombian farmers (Cardenas et al., 2000) had the form of a common pool problem. In absence of explicit incentives, extraction levels were not far above the optimum. But when monitoring of extraction levels and a fine for over-extraction were introduced, farmers started to extract more rather than less. Frey and Oberholzer-Gee

(1997) find that people are less likely to accept siting of waste facilities in their neighborhood when they are offered financial compensation for it. Ostmann (1998) provides experimental 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.

In contrast to the growing empirical literature, there is not yet much theoretical work on the informational role of incentives. Two papers model the adverse effects of incentives through their effect on agent's beliefs in a principal-agent context. In Benabou 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.

Sliwka (2007) also considers a principal-agent context, in which there are three types of agents in an organization: altruists, who take into account the principals payoff, egoists, who maximize their own payoff, and conformists, who 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, high incentives signal to the conformists that most people are selfish and this in turn will cause them to exert minimum effort. The principal may thus choose to trust rather than control the agents.

In contrast to these approaches we study a collective action framework. That is, we focus on the impact of incentives on the interaction between agents rather than between agent and principal. We adapt the framework of Benabou and Tirole (2003) by incorporating multiple agents and model strategic interaction between them. As in Benabou and Tirole (2003), the incentives that the government uses carry information, but instead of learning something about their own type, the agents learn something about the type of the other player. In contrast to Sliwka (2007), we do not need to model a psychological game to accomplish a signalling effect of incentives. In our model beliefs do not induce a preference change but serve the more traditional role of anticipating payoffs. The signalling effect arises simply from adding preference heterogeneity to a standard collective action framework.

Finally, outside a principal agent context, Zwart (2006) provides a model where policies designed to solve a problem may backfire because of their effects on the beliefs of agents in the economy. Zwart considers IMF loan provision to a country, where investors have a noisy signal of the economy's fundamentals. Intervention by the IMF signals that the country faces a problem, which may cause the investors to run.

### 3 The model

The model is a sequential game of costly signalling with three different kinds of players: Two 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 (see section 5), but throughout this section we will frame the problem as a public one, and use the words 'government', 'citizens' and 'society'. We focus on pure strategy play.

The central idea is the following: The two citizens play a game of incomplete information. In contrast to standard assumptions, some of the citizens are conditional cooperators. Whether mutual cooperation can be an equilibrium thus depends on the types of the players, which are drawn independently from a distribution determined by nature. Nature thus transforms the game into one of imperfect information. The citizens don't know the distribution of types, but have a prior over the type of the other player.

The government observes the distribution of types in the economy and has an interest in effecting the outcome that yields most utility, i.e. mutual cooperation. It can influence the outcome of the game by introducing what we call sanctions. 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 will make inferences from the sanctions they observe about the type of others 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 appendix C we analyze the role of asymmetric information in the model and show that it causes low sanctions to be played more likely in equilibrium. Furthermore, there is only a partial conflict of interest between the government and the citizens: the government would prefer that the agents knew the distribution of types. However, low types benefit from the fact that their type is only privately known.

## 3.1 The players

### 3.1.1 Nature

Nature moves twice. At the beginning of the game, nature draws a distribution of types for the economy from a class of simple distributions. Every distribution is characterized by the probability  $\rho \in [0, 1]$  that a high type is drawn from it. We call the distribution characterized by  $\rho$  *the state of society*, because it reflects the proportion of conditional cooperators relative to egoists in the economy. The probability that nature picks a given state of society is given by a uniform distribution with support on  $[0, 1]$ .

After nature has chosen the state of society, the government observes it and sets its policy. The second move of nature consists of drawing the two types from the state of society that are matched in a collective action problem described below. The probability that two high types will be matched is thus  $\rho^2$ .

### 3.1.2 The government

The government observes the state of society and therefore the joint distribution of the two types that will be matched. On the basis of this information it sets incentives  $g \in \mathbb{R}^+$ . The objective is to induce cooperation by the citizens in the economy. The instrument to do so is the use of 'incentives', a sanction on defection by the agents. (We will use the words 'sanctions' and 'incentives' interchangeably.) However, introducing such incentives comes at a cost. The principal's objective function is:  $W = 1_{(C,C)}B - \lambda g$ .

Here,  $B$  is the payoff the principal receives from cooperation of the agents, and the operator  $1_{(C,C)}$  takes the value of 1 when both agents in the economy cooperate and 0 otherwise. The principal is thus exclusively motivated by effecting the cooperative outcome. This can be interpreted as the the social benefit that society experiences from a cooperative outcome, or the profit that a principal makes if his employees work together. Since many applications of this result are in the public realm, we show in appendix A that all results also hold for a utilitarian welfare function of the form  $W = \sum_i u_i - \lambda g$ .

The costs of setting incentives are given by  $\lambda g$ . We offer two interpretations for the idea that higher sanctions carry higher cost. First, one can interpret them as the practical costs of setting higher sanctions, such as putting police on the street, building jails, maintaining a more extensive judiciary or raising the probability of getting caught<sup>1</sup>. Second, the parameter  $\lambda$  can measure the moral cost of high sanctions: high sanctions are likely to meet resistance based on the idea that the punishment should not exceed the crime. Although many people would agree that stealing a bike is wrong, few would want to institute the death penalty for bike thieves, even if this were the most efficient way to deter them.

---

<sup>1</sup>One can also interpret  $g$  as the expected cost  $\mu f$  of imperfectly enforced sanctions, where  $\mu$  is the probability of getting caught for defection and  $f$  the level of a fine. The payoff of defection to the agent is  $\pi_D - g$  which can be seen as a reduced form of  $\mu * (\pi_D - f) + (1 - \mu) * \pi_D$ . Then,  $g = \mu f$ , and the term incorporates both the level of a fine and the probability of getting caught. This suggest a natural interpretation of why the cost of setting incentives increases in the size of  $g$ : raising the probability  $\mu$  of catching defectors is increasingly expensive.

Note that we 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, we want to focus purely on the deterring or Hobbesian effect of sanction and not on the revenue-raising aspect. Note also that sanctions (and therefore their costs) are determined in place before citizens decide on their actions. This implicitly assumes commitment by the government to carry out the sanctions once they are in place. We think this is natural in a setting where sanctions are decided upon by politicians, and their execution and enforcement then carried out by the executive and judiciary branch of government.

### 3.1.3 The citizens

Nature randomly draws two citizens  $i$  and  $j$  from the state of society, who have to solve a collective action problem. That is, they play a simultaneous move game of incomplete information in which they have to choose a pure strategy  $s_{i,j} \in \{C, D\}$ , which specifies whether they cooperate or defect. The citizen's utility from cooperation and defection is given by:

$$u_i(C, s_j) = \pi(C, s_j) + \theta_i 1_{(C,C)} \quad (1)$$

$$u_i(D, s_j, g) = \pi(D, s_j) - g \quad (2)$$

Here, the function

$$\pi : \{C, D\} \times \{C, D\} \rightarrow \mathbb{R}^+ \quad (3)$$

maps the two strategies into a positive payoff. The two additive terms in (1) and (2) are not standard and warrant further justification. From (2) we see that the payoff from defection is diminished by the sanction  $g$  set by the government. In (1), the operator  $1_{(C,C)}$  was already defined above and has value 1 if  $s_i = s_j = C$  and 0 otherwise. Therefore, citizens derive utility from their payoff as usual, but also value cooperation as an outcome in itself. The parameter  $\theta_i \in \{0, 1\}$  indicates the importance of this additional payoff to the individual citizen. A low type ( $\theta = 0$ ), or egoist, corresponds to the standard agent in economic literature that cares only about his own payoffs. However, a high type ( $\theta = 1$ ) will behave like a conditional cooperator as will be explained below. Nature draws the parameter  $\theta$  from the state of society before the citizens choose their action. The citizens know their own type but not that of the other player.

Before we give a general structure of the payoffs in the game, we illustrate the framework with an example. Consider the following payoff matrix:

	$C$	$D$
$C$	$4 + \theta_i, 4 + \theta_j$	$0, 5 - g$
$D$	$5 - g, 0$	$1 - g, 1 - g$

The characterization of the game expressed in this matrix depends on both the types and the principal's policy. For  $\theta_{i,j} = 0$  and a  $g < 1$ , the game is a prisoners dilemma. But for  $\theta_{i,j} = 1$  the game is not a prisoners' dilemma but a coordination game, due to the additional payoff attached to the cooperative outcome by high types. Furthermore, for  $g > 1$  cooperation is a dominant strategy for all types.

Note that in this example a high type will act as a conditional cooperator. To see this, we define  $p_i$  as the probability that citizen  $i$  attaches to the event that citizen  $j$  will cooperate. If we compare the expected utility from each action, we see that for a high type,  $E[u_i(C)] \geq E[u_i(D)]$  implies  $p_i \geq 1 - g$ . That is, she will cooperate if and only if her belief that the other citizen cooperates is high enough (relative to the sanction). This is exactly the definition of a conditional cooperator as defined in the introduction.

Embedding this example in a more general notation we have:

$$\begin{aligned} E[u_i(C)] &\geq E[u_i(D)] \\ p_i[\pi(C, C) + p_i\theta_i] + [1 - p_i]\pi(C, D) &\geq p_i[\pi(D, C) - g] + [1 - p_i][\pi(D, D) - g] \end{aligned}$$

Rearranging yields:

$$p_i \theta_i \geq p_i [\pi(D, C) - \pi(C, C)] + [1 - p_i] [\pi(D, D) - \pi(C, D)] - g$$

Since we want to study collective action problems we assume that in absence of preferences for cooperation and sanctions, defection is a dominant strategy :  $\pi(D, C) - \pi(C, C) = \pi(D, D) - \pi(C, D) = 1$ . Moreover, we assume that mutual cooperation Pareto dominates mutual defection:  $\pi(C, C) > \pi(D, D)$ . As in the example above, this makes the characterization of the game dependent on the type of the citizens. Abstracting again from the government policy, we see that for low types the game is a prisoners' dilemma. For high types it is a coordination game. Moreover, we have an interaction between sanctions and beliefs, because a citizen  $i$  will cooperate if and only if:

$$p_i \theta_i > 1 - g \tag{4}$$

For simplicity we will adopt the tiebreak-rule that indifferent people cooperate<sup>2</sup>. Then, (4) tells us that a low type ( $\theta = 0$ ) will only cooperate when  $g \geq 1$ . In other words, egoists have to be forced to cooperate by high sanctions. High types ( $\theta = 1$ ) on the other hand, will cooperate whenever  $p_i \geq 1 - g$ . That is, they behave like conditional cooperators. Note that in this setup a reward for cooperation would have exactly the same effect as a sanction on defection. The important thing for our model to work is that the government raises the expected utility of cooperation of the citizens at a cost to itself.

We call "trust" the belief of a player that the other player is a high type. Note that a certain amount of trust defined in this way, is a necessary condition for a conditional cooperator to cooperate if sanctions are low ( $g \leq 1$ ). However, it is not a sufficient one. Believing that the other is a high type does not imply that one believes that the other will actually cooperate. Two conditional cooperators truly face a coordination game. Therefore, our model will not be able to distinguish between multiple equilibria that are associated with trust: one where conditional cooperators are able to coordinate on cooperation, and one where they are not.

In sum, the citizens are characterized by two functions. The first specifies belief generation. Its domain is the own type and government policy. Its range is the belief about the type of the other citizen:  $b : \{0, 1\} \times \mathbb{R}^+ \rightarrow [0, 1]$ . The second function specifies the choice of strategy. Its domain is over the citizens own type and the government policy. Its range is the strategy space:  $s : \{0, 1\} \times \mathbb{R}^+ \rightarrow \{C, D\}$ .

### 3.2 Timing of the game

Reiterating, the timing of the game is as follows:

1. Nature chooses the state of society characterized by the probability  $\rho$  that a high type is drawn from it.
2. The government observes  $\rho$  and decides on its policy  $g$ .
3. Nature draws two citizens from the state of society who are matched to play a collective action problem.
4. The citizens learn their own type and the government policy  $g$ , update their prior, and choose their strategy  $s \in \{C, D\}$ .

---

<sup>2</sup>Alternatively, one could raise all the equilibrium sanctions in the paper by  $\epsilon$ .

## 4 Crowding out of trust

In this section we characterize the perfect Bayesian equilibrium of the game, consisting of government and citizens strategies and citizens beliefs:  $(g, s_i, s_j, b_i, b_j)$ . After, we discuss its implications. Before doing so, it is useful to develop some terminology. We refer to the government policy in a pooling equilibrium simply as  $g$ , since the principal always sets the same sanction. We will see that there is what we call a partial pooling (or semi-separating) equilibrium: an equilibrium with two regions or steps in each of which the government plays the same policy. The boundary value of  $\rho$  between the regions is called  $x$ . We call a region where  $\rho \in [0, x)$ , i.e. where society consists of relatively many egoists a ‘bad state of society’ and those where  $\rho \in [x, 1]$  a ‘good state of society’. We label the government policy for this semi-separating equilibrium as follows: The policy that is set for the bad state of society is called  $g_1$  and the policy for the good state of society is called  $g_2$ . Finally, for our proofs we need to define off-equilibrium beliefs  $b_{oe}$  as the beliefs that citizens have when they see a policy not observed in equilibrium.

We first establish that there is crowding out of beliefs in equilibrium:

**Proposition 1** *If setting sanctions is not too costly relative to the payoff from cooperation ( $\frac{B}{\lambda} \geq 1$ ), then there is a unique pure strategy equilibrium in which high types coordinate on cooperation. This equilibrium features pooling in two intervals, and is characterized by higher sanctions in the bad state of society:  $g_1^* > g_2^*$ . In this equilibrium the government sets  $g_1^* = 1$  to coerce everyone to cooperate, and it sets  $g_2^*$  such that it induces cooperation only from the high types.*

**Proof.** There are at most three different policies in equilibrium. The reason is that there are at most three different situation in the economy: One where everybody cooperates, one where only the high types cooperate, and one where nobody cooperates. If there were more than three policies, two policies must correspond to the same situation. This cannot be an equilibrium since the government would always deviate to the cheaper policy. This means that the three policies that are equilibrium candidates are the ones that most cheaply induce the three situations described above. From the payoffs in the citizens’ subgame, we see that setting  $g = 0$  and  $g = 1$  is the cheapest way of having respectively no cooperation and full cooperation. Whenever  $\frac{B}{\lambda} \geq 1$  we have that setting  $g = 1$  and inducing full cooperation always yields a higher payoff than setting  $g = 0$  and leaving everybody to defect, therefore, setting a sanction of 0 cannot be part of the equilibrium policy.

Now we can get the result by eliminating all other equilibria. We start by ruling out pooling equilibria. Suppose pooling on  $g^* = 1$ . For a high type off-equilibrium beliefs  $b_{oe}$  satisfy  $b_{oe} > 0$ , because high types know that the other citizen is drawn from the same distribution as they are. Then there is an optimal deviation at  $\rho = 1$ , because in this case everybody is a high type and they will cooperate for any sanction  $1 - b_{oe} \leq g < 1$ . This equilibrium can therefore not exist.

Suppose pooling on  $g^* < 1$ . In this case there is a deviation to  $g = 1$  at  $\rho = 0$ , because nobody is a high type and there will be zero cooperation under  $g < 1$ . Note that the fact that the government can coerce citizens means that pessimistic off-equilibrium beliefs are not enough to support such a pooling equilibrium. We thus have a two step semi-separating equilibrium. We know that  $g_1^* = 1$  because any  $g_1 < 1$  yields a deviation at  $\rho = 0$ . We know that  $g_2^* \neq 0$  because setting  $g_2 = 0$  yields no cooperation and a zero payoff. It follows that  $g_2$  is such that it is the cheapest way to induce the high types to cooperate (characterized below). However, for there to be no deviation to  $g_2 < g_2^*$  we need that off-equilibrium beliefs are more pessimistic than those induced by  $g_2^*$ . ■

Proposition 1 is the main result of this paper, and says two things. First, when there are many conditional cooperators, it is best to implement low sanctions and tolerate a few defectors. This is a simple consequence from the existence of conditional cooperators, but often overlooked in discourses on collective action. Second,

there is crowding out of trust in equilibrium, because higher sanctions are associated with a bad state of society. The intuition is straightforward: a government will punish heavily when she knows that most likely there will be a lot of egoists around, because this is the only way to insure substantial amounts of cooperation in such an environment. It will punish less heavily when it expects citizens to be conditional cooperators, because cooperation can be induced cheaply in such an environment by setting lower sanctions.

What drives this crowding out result? First, society must be a mixture of different agents. Second, setting higher sanctions must be costly to the government. If incentives are free to administer, the government would always put a perfectly enforced death penalty on every defection and no signal could be provided. The fact that the government knows  $\rho$  and the agents do not, also influences the threshold value  $x$ . In fact, relative to a situation with complete information about  $\rho$ , the asymmetric information lowers  $x$  and enlarges the region where the low fine is played. The reason is that the government can induce or 'crowd in' trust of citizens by playing a low fine. This is explained in more detail in section 5.

The reason we can rule out equilibria that feature pooling on low sanctions supported by pessimistic off-equilibrium beliefs, is that for bad states of society the government would deviate to  $g = 1$ . In this case the government coerces everybody to cooperate, and this makes the off-equilibrium beliefs irrelevant. We can rule out the pooling equilibrium on  $g = 1$  as well, because in a perfect world where everybody is a high type, the government can induce full cooperation for a sanction lower than one.

The conditions for the crowding out equilibrium to exist are intuitive to understand. First, the high types need to coordinate on the cooperative outcome. That is, if a high type thinks the other player is a high type, she also thinks the other will cooperate. There is also a 'Hobbesian' pooling equilibrium in which high types coordinate on mutual defection and the government sets  $g = 1$ . This makes high types behaviorally equivalent to egoists, and our analysis collapses to the standard one. The existence of this equilibrium is a consequence of the fact that the model is not able to select between the multiple equilibria in the game between conditional cooperators.

Second, other equilibria arise if the cost of setting sanctions becomes too high relative to the benefits of cooperation ( $\frac{B}{\lambda} < 1$ ), because the coercive sanction  $g = 1$  is now dominated by the low sanction  $g = 0$ . It essentially decides that inducing cooperation is not worth the effort, at least when there are many egoists. Consequently, there is a pooling equilibrium on  $g^* = 0$ , supported by pessimistic off-equilibrium beliefs. There is also an equilibrium that features 'crowding in', i.e.  $g_1^* = 0 < g_2^*$ . Only when there are many high types who will cooperate for low sanctions will it be profitable for the government to set any sanctions at all. The reason we do not focus on these cases is that we want to concern ourselves with collective action problems for which the stakes are relatively high. We think that in the most important collective action problems in society, such as tax compliance, the benefits of cooperation justify the costs of high sanctions.

Third, if high types are very optimistic when they see a sanction that is not played in equilibrium, the government will deviate from the equilibrium in Proposition 1 for all  $\rho > x$  where  $x < x^*$ , because off the equilibrium path it is cheap to induce high types to cooperate.

Proposition 2 characterizes the equilibrium in Proposition 1 further:

**Proposition 2** *The equilibrium described in Proposition 1 has the following characteristics:*

- a) *the threshold distribution  $x$  is increasing in  $\frac{B}{\lambda}$ . Thus, when the relative cost of setting sanctions goes up, the high sanction is less likely to be observed,*
- b)  *$g_2$  is decreasing in  $\frac{B}{\lambda}$ . When setting sanctions becomes relatively more expensive, the sanctions in the good state of society increase.*

**Proof.** The proof proceeds in a few steps. First we characterize the agents' belief about the distribution of types in the economy. Agents base their beliefs on the principal's policy and their own type. We are interested

only in the belief of conditional cooperators (high types) under a sanction  $g < 1$ , because this is the only case in which beliefs matter for the choice of action. Therefore, we look at the belief of agents that observe  $g_2 < 1$  and  $\theta = 1$ . The common prior is that each distribution is equally likely to be chosen by nature (and consequently the probability that the other is a high type is  $\frac{1}{2}$ ). Conditional on these pieces of information we compute the posterior belief  $b(\rho)$  that a given distribution  $\rho$  has been chosen by nature from Bayes' rule:

$$\begin{aligned}
b(\rho) &= P(p = \rho \mid \rho \geq x, \theta = 1) \\
&= \frac{P(p = \rho \cup \rho \geq x \cup \theta = 1)}{P(\rho \geq x \cup \theta = 1)} \\
&= \frac{\rho\left(\frac{1}{1-x}\right)(1-x)}{\left(\frac{x+1}{2}\right)(1-x)} \\
&= \frac{2\rho}{1-x^2}
\end{aligned} \tag{5}$$

This belief is increasing in  $\rho$ : a high type is more optimistic about society than a low type (for which the belief would be decreasing in  $\rho$ ).

Second, we look at the best response of the citizens in the economy to any government policy given their belief 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 under  $g_2$ . The best response of a high type is to cooperate when his expected utility from cooperation is positive. Expected utility depends on  $b(\rho)$  and  $g$ . A conditional cooperator knows that in equilibrium the other player will either cooperate for sure (under  $g_1$ ) or cooperate only if she is a high type (under  $g_2$ ). For each value of  $x$  there is a sanction  $g_2$  such that the high type exactly cooperates given her beliefs. The calculation is provided in appendix B, and yields the relation:

$$g_2(x) = 1 - \frac{2(1-x^3)}{3(1-x^2)} \tag{6}$$

We can verify that  $g_2$  is decreasing in  $x$  which is intuitive: When  $x$  increases, the environment in which the principal plays  $g_2$  shrinks. Therefore, high types are more optimistic when  $g_2$  is observed and this in turn lowers the sanctions needed to induce the high types to cooperate.

Third, now that we know the reactions of all the agents in the economy to all possible government strategies, we just need to find the government's optimal strategy. The government will set sanctions  $g_2$  according to (6), because it is the cheapest way to induce cooperation from the high type. Since we know that  $g_1 = 1$  and  $g_2$  is determined from (6), we need to find  $x$ , which is the state of society for which the government is indifferent between the two strategies. It is found by setting:

$$E[W(g_1 = 1)] = E[W(g_2)]$$

At distribution  $x$ ,  $g_1$  yields the cooperative outcome for sure and  $g_2$  yields the cooperative outcome when both types are high, that is, with probability  $x^2$ . We can therefore write:

$$B - \lambda = x^2 B - \lambda g_2(x)$$

Substituting from (6) we find:

$$\frac{B}{\lambda} = \frac{2(1-x^3)}{3(1-x^2)^2} \tag{7}$$

With the help of the implicit function theorem we can establish that  $x$  is increasing in  $\frac{B}{\lambda}$ . This establishes part a). From (6) we know that  $g_2$  is decreasing in  $x$  which establishes part b). ■

Figure 1 illustrates Proposition 2a). The uninterrupted lines delineate an initial situation that results in a threshold  $x$  (the point where the expected payoffs are the same). The policy  $g_1$  yields constant welfare of  $B - \lambda$ ,

because all types cooperate. We see that the welfare from policy  $g_2$  increases in the state of society, because it depends on the proportion of high types. It increases quadratically in  $\rho$ , because the probability that *both* agents are high types is  $\rho^2$ .

Suppose  $B$  falls exogenously to  $B'$ . We can think of this as a fall in the political priority of cooperation in the collective action problem under consideration. As a result the relative benefit of cooperation  $\frac{B}{\lambda}$  decreases, expected payoffs from both policies fall, but the more so under the high sanction because this gives the principal benefit  $B$  for sure. As a consequence,  $x$  shifts to the left. That is, the region of distributions for which the principal will set  $g = 1$  shrinks. In other words, if the benefits of cooperation decline, a government is less anxious to ensure cooperation through coercion. Similarly, if  $\lambda$ , the cost of setting sanctions goes up,  $\frac{B}{\lambda}$  falls, and higher sanctions will be played less likely.

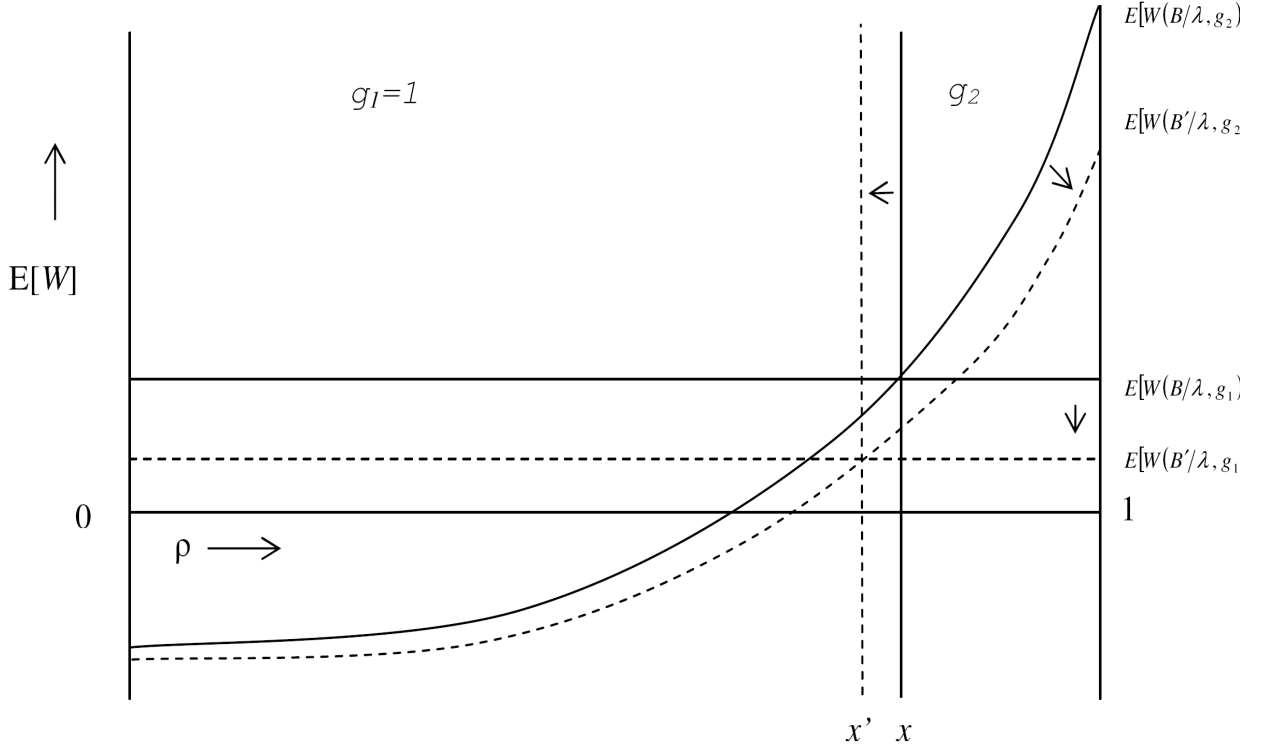


Figure 1: An exogenous decrease in the relative benefit of cooperation.

Note that, no matter how high  $\frac{B}{\lambda}$  becomes, high sanctions will never be sure to be played ( $x$  will never quite reach 1). The reason is that there will always be a profitable deviation to a lower penalty with a perfectly informative signal at  $\rho = 1$ . Also, from (7) we can compute that in the crowding out equilibrium,  $x$  will never be smaller than 0,48. For lower values of  $x$  we need that  $\frac{B}{\lambda}$  falls below 1, but as we have seen this case leads us into a new equilibrium.

Result 2b says that when the relative cost of sanctions increases, paradoxically, the sanctions go up for the good state of society. The reason for this result is that when  $\frac{B}{\lambda}$  decreases the principal will enlarge the region where it plays  $g_2$ . This means citizens adjust their beliefs upon seeing  $g_2$  downwards. Consequently, the principal will have to set a higher sanction to get the high types to cooperate.

In sum then, the results in this section show that under the conditions mentioned above, higher sanctions may diminish trust in society. The reason is that the principal sets high sanctions only if there are a lot of egoists, which discourages the high types. Appendix A shows that the results in this section hold similarly for a utilitarian social welfare function.

## 5 The role of asymmetric information

An important aspect of the model is the asymmetric information about the state of society. Here, we compare the case of asymmetric information with a case of complete information, i.e. in which the citizens would know  $\rho$ . The derivations of the results in this section are provided in appendix C.

The most important implication of asymmetric information, is that a government can ‘crowd in’ cooperation by the citizens. That is, the fact that the citizens do not know  $\rho$  causes low sanctions to be played more frequently, by shifting down  $x$  relative to a world where citizens know  $\rho$ . Therefore, low sanctions are more likely under asymmetric information.

To see this, consider the case when there is complete information about  $\rho$ . Similar to the asymmetric information case, there will be a region of bad states of society in which government sets  $g_1 = 1$ . There is again a threshold state of society that we call  $d$ , where the government finds that coercion is too expensive relative to relying on trust between the conditional cooperators. As before,  $d$  depends on  $\frac{B}{\lambda}$ . When the state of society is above the threshold  $d$ , the government will then set  $g = 1 - \rho$ , and the high types will cooperate.

To compare the two cases we compare  $x$  and  $d$ . However, for lack of explicit expressions, we have resort to simulations. The result is illustrated in figure 2. We see that  $x < d$ , which means that asymmetric information enlarges the set of states of society where low sanctions are optimal.

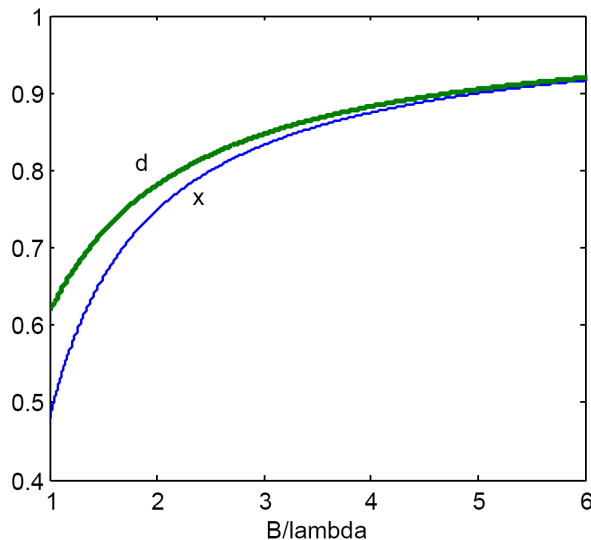


Figure 2: Threshold policies under symmetric ( $d$ ) and asymmetric ( $x$ ) information.

The intuition behind this result is that the government induces trust of citizens by setting a low sanction. This is possible because the citizens are unsure about the state of society. Consider the lowest possible distribution  $d$  for which the principal will still set low sanctions under symmetric information. Here, beliefs will be  $d$  under symmetric information, because the citizens know exactly what’s going on. However, under asymmetric information beliefs will be  $b_i(g, \theta_i)$ . Setting low sanctions in this case induces beliefs  $b_i(g, \theta_i) > d$  and enlarges range in which he sets a low fine under asymmetric information. An implication of this result is that, paradoxically, when information in society about the behavior of others becomes more accurate, the government will on average (taken over all states of society  $\rho$ ) set higher sanctions.

However, it is that the ex-ante (before nature chooses  $\rho$ ) welfare of the government would be higher if people knew the state of society. Therefore, we should not conceive the signal provided by incentives as revealing

something that the government would like to cover up. Rather, the government would (at least ex-ante) prefer  $\rho$  to be public, but can not credibly make it so. The signalling 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.

The reason for this result is that although the government benefits from the crowding in described above, it cannot benefit from the optimism in that would exist in extremely good states of society in a situation of full information. In the latter situation, it would be able to set even lower sanctions than it can in the case of asymmetric information.

In sum, asymmetric information enlarges the region where low sanctions are played, and thereby lowers average sanctions. However, the government is ex-ante better off under symmetric information.

## 6 Implications and discussion

We start the discussion with some general remarks, and continue by presenting two extensions to the model that highlight the importance for policy making of taking into account heterogeneity of types.

First, there is a role for government even in a society that consists mainly of conditional cooperators, where the collective action problem looks like a coordination problem rather than a prisoners dilemma. Although their behavior is largely driven by trust, conditional cooperators will still need a little 'push in the back', because they are aware that there are some egoists around which reduces their desire to cooperate. With respect to the egoists, tolerating the few defectors that there are may have a lower price tag than introducing costly high sanctions. This latter result follows directly from the existence of conditional cooperators, but is nevertheless overlooked in many discourses on collective action. Note that this result does not depend on signalling or on asymmetric information: if citizens perfectly knew the state of society, it would still be true.

Second, the result does not say that compliance goes down when sanctions go up. As in Benabou and Tirole (2003), incentives are what they call 'short term reinforcers'. The crowding out does not occur on the level of behavior, because in this model the higher sanctions 'override' the effect of diminished beliefs. That is, under a high sanction, people in society are coerced into cooperation, but will think of their peers as being essentially egoists. Thus, an econometrician looking solely at the relation between sanctions and cooperation, would support the standard Becker-Stigler results. However, the econometrician might not observe that as a result of high sanctions, trust in society is diminished.

Third, the model can explain why a heightened political awareness of a certain collective action problem can result in increased sanctions. We can interpret a rise in  $B$  as an increase in the political importance of a collective action problem. For example, a new government with different priorities can be voted in office. As proposition 2a tells us, the threshold value  $x$  between the two policies increases in this case because it now becomes more important to ensure full compliance. If the political change causes  $x$  to increase above the true state of society, sanctions will increase. The model predicts that such a change in policy will increase compliance, but will reduce trust. Interestingly, if the true state of society remains above  $x$  even after the increase in  $B$  then sanctions will actually go down. The intuition here is that conditional cooperators become more optimistic when they see that even under the increased political urgency of cooperation no high sanctions are necessary.

### 6.1 Decreased trust and addiction to sanctions

A way in which decreased trust can have behavioral consequences is through spillover effects into the future and into other collective action problems. Trust is an attitude that determines behavior in many social situations. The crowding out of trust by incentives in one area could therefore have spill-over effects in other areas. Consider

a number of similar collective action problems. In each of them, a citizen is randomly matched with another citizen, whose type is drawn from the same state of society always. Suppose that the government considers policy for only a single collective action problem, and sets  $g = 1$ . This will crowd out trust and the citizen may now defect in the other collective action problems where sanctions are not coercive enough to overrule the crowding out effect. Thus, a raise in sanctions in one policy area may cause a drop in cooperative behavior in other areas.

Secondly, as remarked by Benabou and Tirole (2003), one can imagine a situation where people think they would be able to get away with defection, for example because of imperfect monitoring. In this case, only the negative signalling effect remains, whereas the coercive effect of incentives disappears. Another example of this is when one group in society is subjected to higher sanctions, and another is not. The group that is exempted may form a more negative opinion of the sanctioned group, even though the people in this group may now cooperate more. Cooperation of the non-sanctioned with the sanctioned will go down. Section 5 shows elaborates this idea in an application to support for the welfare state.

Finally, sanctions may 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 exogenously lowered (for reasons not described in the model) after they have been introduced, cooperation may see a large drop, as even the by now cynical high types will refuse to cooperate. This is consistent with experimental evidence as in Gneezy and Rustichini (2000) or Gächter (2007). These studies show that when incentives are withdrawn, cooperation does not return to pre-incentive levels.

A proper analysis of these cases is a task for future research. Suvorov (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 resort to ever more controlling measures to compensate for the reduced trust.

## 6.2 The welfare cost of treating people as egoists

The standard view espoused by much of modern economics and political science holds that people are by and large egoistic, i.e.  $\rho = 0$ . Given the prominence of this paradigm, we briefly look at the social welfare resulting from a ‘Hobbesian’ government that embraces this view. That is, suppose that the government does not maximize its objective function taking into account its information, but instead sets a blanket policy of high incentives. Depending on the benefits of cooperation and the cost of setting sanctions, this may be inefficient in terms of the government objective function, because most people can be induced to cooperate by convincing them that others do so, without recourse to costly sanctions.

The ex-post (after nature has chosen the state of society) efficiency cost  $C$  of setting the high sanctions in a good state of society is:

$$\begin{aligned}
 C &= E[W(g_2)] - E[W(g_1)] \\
 &= (\rho^2 B - \lambda g_2) - (B - \lambda) \\
 &= B(\rho^2 - 1) - \lambda(1 - g_2) \\
 &= B(\rho^2 - 1) - \lambda \frac{2(1 - x^3)}{3(1 - x^2)} \\
 &= B(\rho^2 - x^2)
 \end{aligned} \tag{8}$$

Where we used (6) to go from the third to the fourth line and (7) to go from the fourth to the fifth line. Note that  $x = x(\frac{B}{\lambda})$  so that the cost of misspecifying the penalty is a function only of  $\frac{B}{\lambda}$  and the state of society  $\rho$ .

The intuition underlying (8) is straight forward. Costs increase in the distance of the realized state of society from the policy threshold  $x$ , i.e. when there are more conditional cooperators in society. The further this distance, the bigger the 'mistake' in setting high sanctions. Furthermore, costs increase in  $\lambda$ , because a higher  $\lambda$  lowers  $x$ . The effect of  $B$  is ambiguous: On the one hand, cost rise in  $B$  because more money is at stake, so a higher  $B$  amplifies any error of the policy maker. On the other hand, a higher  $B$  raises  $x$  and thereby lowers cost.

## 7 Applications

The potential applications of the model described in this paper are various. In fact, they 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. We discuss two applications that are very important in modern societies for which we think our model is relevant: tax compliance and the support for the welfare state. Other applications in the public realm include fare evasion and not in my back yard (NIMBY) problems. In the context of organizations and personnel economics, one can apply the model to team work. A forceful and more extensive argument for the application to tax-evasion as well as NIMBY and crime can be found in Kahan (2005). In all of these cases there is circumstantial evidence that the adverse effect of incentives outlined above is at work. Isolating and assessing the economic importance of these effects could be the subject of future research.

### 7.1 Tax compliance

Tax evasion fits the model well, because it is a hidden activity: Any single tax-payer has very incomplete information on how honestly others pay their taxes. Tax offices on the other hand collect statistics on evasion rates. This makes tax-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 show that both on an individual level (Scholz, 1998) and on an aggregate (national) level (Frey and Torgler, 2006) the decision to evade taxes is largely based on dispositional attitudes. Especially important are the belief that fellow taxpayers evade and the perceived legitimacy of the use of tax revenue.

Experimental evidence yields the same results. Coleman (1997) reports the results of the Minnesota tax experiment, in which 47,000 tax payers received a letter announcing increased audit probability. The responses with respect to reported income were mixed. In one treatment, the experimenters sent another letter to 20,000 tax-payers saying that the numbers of cheating tax-payers 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 publicity about these campaigns may increase distrust in other citizens. This distrust in turn is a determinant of evasion in their estimations.

Furthermore, this paper can explain the fact that fines and audit rates are generally too low to be credible deterrents. Instead of relying on deterrence, governments realize that people are motivated by reciprocal preferences and choose to apply a low fine. Another prediction of the model is that in equilibrium, high sanctions on tax evasion only make a difference for low types. High types will pay their taxes for any equilibrium sanction. This fits well with empirical results. 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.

## 7.2 Support for the welfare state

Conditional cooperation is one of the driver forces behind the welfare state. Fong et al. (2005) show based on a multitude of evidence that the most important reason for people to oppose the welfare state is the conviction that the poor are lazy. This is true for the poor as well as the rich. By contrast, people indicate in interviews that they are willing to support those that are temporarily unlucky but willing to search for a job. This research is conducted in the United States, but the findings fit perfectly with those of Mau (2003) who compares support for the welfare state in both Germany and the United Kingdom<sup>3</sup>.

In such a setting, the model predicts that strong sanctions on laziness of the unemployed (i.e. unwillingness to look for a job) will encourage job search and reduce benefit-dependence through the coercive effect. Paradoxically however, it will also reinforce the idea among employed and unemployed that the unemployed are lazy. Because the employed are not subjected to sanctions that force them to cooperate, for them the crowding out effect will dominate and undermine their willingness to cooperate with the unemployed, i.e. support the welfare state.

## 8 Concluding remarks

This paper asks whether Hobbesian coercion in collective action problems remains optimal when society is a mix of conditional cooperators and egoists. What is the optimal policy to promote cooperation if the collective action problem in question is a prisoners dilemma for some and a coordination game for others? The paper shows that the optimal level of sanctions depends on the relative proportions of the two agents in society and describes how official incentives can crowd out trust in rational agents. When there are many egoists, the high sanction or Hobbesian solution is optimal. When there are many conditional cooperators, a policy of low sanctions is more efficient. This means that high sanctions crowd out, and low sanctions crowd in trust. This in turn implies that its superior information allows government to induce or crowd in cooperation by setting low sanctions. It will therefore set lower sanctions on average relative to a situation with full information.

There are three main conditions that drive this crowding out result: the citizens are a heterogenous mixture of conditional cooperators and egoists, the government has more information than its citizens about the type of people in society, and sanctions are costly to implement. Under these circumstances sanctions provide the conditional cooperators with a cue about the proportion of egoists, which is crucial for their willingness to cooperate. The paper thus shows that sanctions may have a dual role. They both change economic payoffs and alter agents' perception of the environment.

When the environment consists predominantly of conditional cooperators, a picture that seems to emerge from experimental evidence, cooperation can be mainly achieved by relying on trust. Under such circumstances, a policy implemented by a Hobbesian minded government that departs from the view that people are egoists is unnecessarily costly. It may also generate negative social effects that are associated with diminished trust.

There are important aspects of collective action that the model does not capture. One such aspect are framing effects. Sanctions could be accompanied by signals that guide their interpretation. A principal that manages to punish the egoists, but at the same time convey the message that many people are in fact complying, will not crowd out trust. In contrast, she will reassure the conditional cooperators that they are not suckers that pay the collective bill on their own. To take up the example of tax evasion again, there is evidence that the nature of publicity about tax evasion matter. Sheffrin and Triest (1992) expose groups in an experimental setting to two statements about tax evasion. One says that the tax authorities have stepped up detection efforts. The other adds that these increased efforts are taken because the 'tax gap' (due taxes that have remained uncollected) has reached 100 billion dollars. The authors show that the second statement significantly decreases trust in

---

<sup>3</sup>For a more extensive survey, see Benabou and Tirole (2006).

other taxpayers. In the econometric part of their paper this latter variable is found to be a determinant of the decision to evade.

Another limitation is the static nature of the model that makes it impossible to evaluate potential forward looking effects of sanctions. To the extent that people expect others to be more likely to cooperate after higher sanctions are introduced, the effects of sanctions on beliefs may be more complicated.

## 9 References

- Bénabou, Roland and Jean Tirole. 2006. "Belief in a just world and redistributive politics." *The Quarterly Journal of Economics*, 121:2 pp. 699-746.
- Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies*, 70, pp. 489-520.
- Cardenas, Juan Camilo, John Stranlund and Cleve Willis. 2000. "Local Environmental Control and Institutional Crowding-Out." *World Development*, 28(10), pp 1719-1733.
- Coleman, Stephen. 1997. "Income tax compliance: A unique experiment in Minnesota". *Government Finance Review*, 13, pp. 11-15.
- Fehr, Ernst and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14:3, pp. 159-81.
- Fischbacher, Urs and Simon Gächter. 2006. "Heterogeneous Social Preferences and the Dynamics of Free Riding in Public Goods." IZA Working Paper 2011.
- Fong, Christina M., Samuel Bowles, and Herbert Gintis. 2005. "Reciprocity and the Welfare State," in *Moral Sentiments and Material Interests*. Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr eds. Cambridge. Massachusetts: MIT Press, pp. 277-302.
- Frey, Bruno S. and Reto Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys*, 15:5, pp. 589-611.
- Frey, Bruno S. and Stephan Meier. 2004. "Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment", *American Economic Review*, 94(5), pp. 1717-22.
- Frey, B. S., & Oberholzer-Gee, F. (1997). "The cost of price incentives: an empirical analysis of motivation crowding-out". *American Economic Review*, 87 (3), pp. 746-755
- Frey, Bruno S. and Benno Torgler. 2007. "Tax Morale and Conditional Cooperation." *Journal of Comparative Economics*, Forthcoming.
- Gächter, Simon. 2006. "Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications." CeDEx Discussion Paper, 2006-03.
- Gächter, Simon, Esther Kessler, and Manfred Königstein. 2007. "Performance Incentives and the Dynamics of Voluntary Cooperation." mimeo.
- Gneezy, Uri and Aldo Rustichini. 2000a. "A Fine is a Price." *Journal of Legal Studies*, 29, pp. 1-17.
- Ichino, Andrea and Maggi, Giovanni, 2000. "Work Environment And Individual Background: Explaining Regional Shirking Differentials In A Large Italian Firm," *Quarterly Journal of Economics*, 115, 1057-1090.
- Kahan, Dan M. 2005. "The Logic of Reciprocity: Trust, Collective Action, and Law," in *Moral Sentiments and Material Interests*. Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr eds. Cambridge, Massachusetts: MIT Press, pp. 339-78.

- Kolm, Ann-Sofie, Hedstrom, Peter and Yvonne Aberg. 2003. "Social Interactions and Unemployment". Uppsala University Economics Working Paper No. 2003:18.
- Mau, Steffen. 2003. *The Moral Economy of Welfare States, Britain and Germany compared*. London: Routledge.
- Ostman, A., 1998. "External control may destroy the commons". *Rationality and Society* 10 (1), pp. 103–122.
- Scholz, John T. 1998. "Trust, Taxes, and Compliance," in *Trust and Governance*. Valerie Braithwaite and Margaret Levi eds. New York: Russell Sage Foundation, pp. 135-65.
- Sheffrin, Steven M. and Robert K. Triest. 1992. "Can Brute Deterrence Backfire? Perceptions and Attitudes in Taxpayer Compliance", in *Why People Pay Taxes*, J. Slemrod ed.
- Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review*, forthcoming.
- Souvorov, A. 2003. Addiction to Rewards, Mimeo GREMAQ, Toulouse.
- Wenzel, Michael. 2004. "The Social Side of Sanctions: Personal and Social Norms as Moderators of Deterrence", *Law and Human Behavior*, 28:5, pp. 547-567.

## 10 Appendix A: utilitarian social welfare<sup>4</sup>

Here I show that the proof of propositions 1 and 2 holds true also if the principal has a utilitarian objective function of the form  $W = \sum_i u_i - \lambda g$ . However, in this case the mechanism in operation is a bit less 'clean' in the following sense: Since high types derive more utility from cooperation than low types, the government's payoff in case the cooperative outcome is reached depends on the types that are around in the economy. The government therefore has a higher motivation to induce the citizens cooperate when they are a high type than when they are a low type. Benabou and Tirole (2003) distinguish this 'profit effect' from the 'trust effect' that arises from the principals' superior knowledge.

We first see under what conditions Proposition 1 holds for the new social welfare function. From the symmetry of the game and our assumptions on the payoffs we can deduce that  $2\pi(C, C) > \pi(C, D) + \pi(D, C)$ . Hence, mutual cooperation gives the highest aggregate utility, and also in this case the government prefers to effect the cooperative outcome. For proposition 1 to hold we need then only find the value of  $\lambda$  for which the government always prefers to set  $g = 1$  rather than  $g = 0$ . For this it is sufficient that:

$$\lambda < 2[\pi(C, C) - \pi(D, D)] \quad (9)$$

Therefore, we can replace condition  $\frac{B}{\lambda} \geq 1$  in Proposition 1 with (9) and the rest of the proof goes through unaltered. This should not be surprising, since the new objective function preserves the essential element needed for the result, namely that payoffs for the government are highest when there is mutual cooperation between the citizens.

With respect to proposition 2, the beliefs and the reactions of the citizens to the government policy are identical. The relation between  $g_2$  and  $x$  is unaltered. The difference comes in the calculation of the optimal government policy. Setting  $E[W(g_1)] = E[W(g_2)]$  now yields the expression:

$$\lambda = 2(1 - x) + 3 \left( \frac{1 - x^2}{1 - x^3} \right) [-x(x + \pi(C, C)) + (\pi(C, C) - \pi(D, D)) + 1] \quad (10)$$

---

<sup>4</sup>Calculations and simulations are available on request.

Again, we can make use of the implicit function theorem to ascertain that  $x(\lambda, \pi(C, C), \pi(D, D))$  is decreasing in  $\lambda$  in the range  $x \in [0, 1]$ . This establishes the equivalence to Proposition 2a) for the utilitarian social welfare function. Proposition 2b) holds because the relation between  $x$  and  $g_2$  is unaltered.

Two further facts are noteworthy. First, we have  $\lim_{\lambda \rightarrow 0} x(\lambda) = 1$ , which also makes intuitive sense: if sanctions are costless, the government always sets the high sanction. Second, simulation yields that  $\frac{\partial x}{\partial \lambda}$  is determined by the values of  $\pi(C, C)$  and  $\pi(D, D)$ . The higher the difference between these two values, the slower  $x$  decreases in  $\lambda$ . This makes sense intuitively: if  $\pi(C, C)$  is high relative to  $\pi(D, D)$ , cooperation yields high welfare and the government will want to make sure that citizens cooperate. Therefore, it will be more likely to set  $g_1$  (i.e. a high  $x$ ) for a given cost.

## 11 Appendix B: best response of high types

A high type that observes  $g_2$  will cooperate only if:

$$\begin{aligned}
E[u(C, s_j)] &\geq E[u(D, s_j, g_2)] \\
\int_x^1 b(\rho) E_\rho[u(C, s_j)] d\rho &\geq \int_x^1 b(\rho) E_\rho[u(D, s_j, g_2)] d\rho \\
\int_x^1 b(\rho) [\rho[\pi(C, C) + 1] + [1 - \rho]\pi(C, D)] d\rho &\geq \int_x^1 b(\rho) [\rho\pi(D, C) + [1 - \rho]\pi(D, D) - g_2] d\rho \\
\int_x^1 b(\rho) [\rho[\pi(C, C) - \pi(D, C)] + \rho] d\rho &\geq \int_x^1 b(\rho) [[1 - \rho][\pi(D, D) - \pi(C, D)] - g_2] d\rho \\
\int_x^1 b(\rho) g_2 d\rho &\geq \int_x^1 b(\rho) [1 - \rho] d\rho \\
g_2 &\geq 1 - \int_x^1 b(\rho) \rho d\rho \\
g_2 &\geq 1 - \frac{2(1 - x^3)}{3(1 - x^2)}
\end{aligned}$$

The government will thus set sanctions exactly in this way, because it is the cheapest way to induce cooperation.

## 12 Appendix C: relaxing asymmetric information<sup>5</sup>

An interesting question how the results depend on the information asymmetry in the model. Note that there is a double information asymmetry: The agent knows more about her own type, whereas the principal knows more about the distribution of types. If the individual types would be known to the principal, he would simply set  $g = 0$  for conditional cooperators and  $g = 1$  for egoists. On the other hand, if the agents knew the distribution of types  $\rho$  the signalling effect would disappear. However, the policy would look similar: The principal would

<sup>5</sup>For want of explicit expressions for  $x$  and  $\delta$ , the results in this appendix rely on simulations. The MATLAB codes for these are available on request.

set  $g_1 = 1$  in bad state of the worlds and  $g_2 = 1 - \rho$  in good state with the threshold value of  $\rho$  between the two policies (call it  $d$ ) depending on the expected payoffs from the two policies. We can compute the relation between  $d$  and  $\frac{B}{\lambda}$ :

$$\begin{aligned} E[W(g_1)] &= E[W(g_2)] \\ B - \lambda &= d^2 B - \lambda(1 - \delta) \\ \frac{B}{\lambda} &= \frac{d}{1 - d^2} \end{aligned}$$

Using the implicit function theorem once more we can establish that  $d$  is increasing in  $\frac{B}{\lambda}$  and  $\lim_{\frac{B}{\lambda} \rightarrow \infty} d = 1$ . Therefore, under symmetric information about the distribution of types the government will set a low sanction only when  $\frac{B}{\lambda}$  is not too high. To see the difference with the asymmetric information case, we compare  $d$  with  $x$ , the lowest observed distribution for which the government sets a low penalty under asymmetric information. For lack of explicit expressions, we simulate  $x$  and  $d$  and we find that  $d > x$  (see figure 2). Under asymmetric information the low sanction is more likely to be played.

Another interesting question is whether the government is better off under asymmetric information. The answer is that it is not. We find this answer by comparing the ex-ante expected welfare of the government in the two situations. Under asymmetric information (*AI*) ex-ante welfare is given by:

$$E[W_{AI}] = (B - \lambda)x - \int_x^1 B\rho^2 - \lambda g_2(x) d\rho$$

Under symmetric information (*SI*) it is given by:

$$E[W_{SI}] = (B - \lambda)d - \int_d^1 B\rho^2 - \lambda(1 - \rho) d\rho$$

Both expressions only depend on  $\frac{B}{\lambda}$  through  $x(\frac{B}{\lambda})$  and  $d(\frac{B}{\lambda})$ . Because we don't have explicit expressions for  $d$  and  $x$  we have to rely on simulations again. These yield that welfare is minimally higher under symmetric information. The government would thus like people to know the state of society, but it can not credibly transfer the information it has.

Finally we want to know whether the expected ex-ante sanction is higher under asymmetric information or complete information. To know this we compare the average sanctions under both regimes that are given by the expressions:

$$E[g_{AI}] = x - (1 - x)g_2(x)$$

Under symmetric information about  $\rho$  it is given by:

$$E[g_{SI}] = d + \frac{1}{2}(1 - d)^2$$

Both expressions only depend on  $\frac{B}{\lambda}$  through  $x(\frac{B}{\lambda})$  and  $d(\frac{B}{\lambda})$ . Once again we have to resort to simulations, and we find that expected sanctions are higher under symmetric information.