Cooperation and social control

Bakker, Dieko Marnix

DOI:
10.33612/diss.98552819

4

Institutional punishment is more robust to oppositional control than peer punishment

Dieko M. Bakker, Jacob Dijkstra, Andreas Flache

This chapter is currently under review at an international peer-reviewed journal

The problem of social control

Communities great and small face problems of social control. Only rarely do members of a group spontaneously all exhibit behavior which is considered desirable by the group as a whole. Often, undesirable behavior is individually rewarding and therefore tempting. Think, for example, of the problem of dog poo on streets. As a society, we prefer to have clean streets. However, picking up dog poo is unpleasant and many dog owners are easily tempted to leave their pet's waste on the sidewalk. What's worse, when such undesirable behavior is not kept under control it has a tendency to spread (Cialdini et al., 1990; Mäs & Opp, 2016). Passers-by observe the presence of dog poo on the sidewalks, infer that many people do not comply with the norm of picking up after your dog and adjust their own compliance accordingly. Similar examples can be found in varied areas of society, from nuisance among neighbors to labor disputes, to the reduction of climate-altering emissions.

A common denominator in these problems is that they can be considered social dilemmas (Kollock, 1998). In social dilemmas, there is an inherent competition between individual and collective interests: individually rational behavior frequently leads to collectively suboptimal outcomes. While the severity of this competition varies (Kollock, 1998; Van Lange et al., 2013) groups frequently need to exercise some measure of social control to enforce collectively desirable behavior.

The problem of social control can be divided into two sub-problems: the problem of monitoring, and the problem of sanctioning (Hechter, 1987). In order to be able to enforce collectively desirable behavior, groups must first be able to monitor the behavior of their members. Only then can undesirable behavior be detected and corrective action be taken. When undesirable behavior is detected, the group must then be able to effectively sanction the deviant group member in order to correct their behavior.

We will set aside the problem of monitoring (for now, see Chapter 5), and focus on the problem of sanctioning. Groups wanting to sanction their members have a broad range of sanctioning systems at their disposal. Apart from variations in the type of sanctions distributed (e.g. fines (Chaudhuri, 2011; Fehr & Gächter, 2002), expressions of disapproval (Andrighetto et al., 2013; Simpson et al., 2017), ostracism (Cinyabuguma, Page, & Putterman, 2005; Masclet & David, 2003; Wiessner, 2005), refusing future cooperation (Axelrod, 1984; Back & Flache, 2008), withholding rewards (Flache et al., 2017)) these sanctioning systems are characterized by their allocation of the power to sanction. In every sanctioning system, there are *controllers*: those individuals or institutions who are allowed to distribute sanctions. This choice of a

controller can be codified, in the way that the jurisdiction of police forces is established by law. Alternatively, as with more informal sanctions such as expressions of disapproval, the choice of a controller can be understood as a social consensus on who can legitimately distribute sanctions (Baldassarri & Grossman, 2011; Ellickson, 2001). Controllers can be, for example, communities of peers, government bodies, organizations or contractual partners.

In the present study, we will compare two sanctioning systems which differ in their choice of controllers. The first of these systems is peer punishment (Fehr & Gächter, 2002; Guala, 2012; Chapter 3 of this dissertation). Under this sanctioning system, all group members are controllers. Each individual member has the authority to distribute sanctions to any of their peers. Translating this to some practical situations, this is similar to informal social control in self-managing teams (Barker, 1993) or to confronting your neighbor when he is playing loud music late at night. Peer punishment systems have been particularly prominent in the literature on social dilemmas (Chaudhuri, 2011; Guala, 2012).

Institutional punishment

As might not be difficult to imagine given these examples, peer punishment is not without its problems and does not appear to be particularly prevalent among social groups outside of laboratory studies (Guala, 2012). Peer punishment may be used to punish collectively beneficial behavior (Herrmann et al., 2008), such as when less productive members of a team chastise high producers for being 'rate busters' (Homans, 1974). Peer punishment can also be vulnerable to retaliation which discourages punishment (Nikiforakis, 2008), such as when confronting a noisy neighbor results in a black eye. Peer punishment is by definition not coordinated or centralized, so both over- and underproduction of sanctions are possible sources of inefficiency in the distribution of sanctions (Bendor & Swistak, 2001; Guala, 2012; Heckathorn, 1990).

As an alternative to peer punishment, we often find institutions which are selected or created to be the controller in a sanctioning system. Police forces, courts, and referees are all examples of such institutions. As controllers, these institutions are given the exclusive right and duty to distribute sanctions to regulate behavior within the boundaries of a social group. The legitimacy and effectiveness of these institutions are dependent on support from the members of this social group. Under democratic regimes, these members also collectively determine what rules are enforced and what sanctions can be distributed. For example, police forces enforce laws which are

created by elected representatives and are dependent on funding from taxes levied on the citizens of the communities in which they operate. Experimental studies have shown that such punishment institutions can be supported by contributions from community members (Yamagishi, 1986) and may be preferred to peer punishment by community members (Traulsen, Rohl, & Milinski, 2012).

Peer punishment and institutional punishment have different costs and benefits. Peer punishment institutions may be cheaper in terms of transaction costs (there is no need to outfit a group of officials, no need for legislation to determine their authority and responsibilities, etc.). However, it suffers from coordination problems, free-rider problems, disagreement among peers on what is and is not acceptable, and so forth. Institutional punishment, meanwhile, may outperform peer punishment in terms of effectiveness (consistent application of the rules, no free-rider problems, higher legitimacy) but has higher transaction costs and does not always account for the different interests of various subgroups.

Previous comparisons between peer punishment systems and institutional punishment systems have focused on aspects such as the legitimacy of punishment (e.g. Baldassarri & Grossman, 2011), the extent of compliance and (antisocial) punishment (e.g. Van Miltenburg, Buskens, Barrera, & Raub, 2014), or sensitivity to unreliable and incomplete information (e.g. Van Miltenburg, Przepiorka, & Buskens, 2017). In this study, we focus on one important cost difference which is particularly relevant to social control: the system's vulnerability to oppositional control (Heckathorn, 1990). As far as we are aware, there are as yet no experimental studies comparing the vulnerability of peer punishment systems and institutional punishment systems to oppositional control. Often studied as retaliation or counter-punishment, a vulnerability to oppositional control is one of the criticisms which have come out in recent years against the effectiveness of peer punishment as a sanctioning system (Guala, 2012; Nikiforakis, 2008).

In this study, we explore the idea that institutional punishment may be less vulnerable to oppositional control than peer punishment, because it is considered to be more legitimate and because it is more difficult to exercise oppositional control against institutional punishment. This may be one of the reasons why we so frequently find institutional punishment systems in social life. We study this experimentally, by implementing peer punishment systems and institutional punishment systems with and without oppositional control, in the context of a 5-person prisoner's dilemma game.

THEORY

Compliant and oppositional control

Sanctioning systems in social dilemmas are intended to enforce collectively beneficial behavior. However, many sanctioning systems also (intentionally or unintentionally) provide opportunities to undermine this enforcement. Heckathorn (1990, 1993, 1996) calls these two behaviors compliant and oppositional control, respectively. Compliant control is directed at non-contributors and is intended to increase contributions. Oppositional control is directed at those who distributed compliant control and is intended to weaken compliant control. In Heckathorn's (1993) interpretation, compliant control is exercised by forcibly reversing another group member's decision not to contribute to a public good, and oppositional control entails blocking another group member's attempt to exercise compliant control. In a broader interpretation, actions intended to make other group members behave according to a certain norm can be seen as compliant control, no matter whether this implies preventing uncooperative behavior or whether it implies applying sanctions after the fact. Similarly, oppositional control is any action intended to prevent compliant control, no matter whether this is done by nullifying the effects of compliant control or whether this is done by discouraging others from exercising compliant control. Closely related concepts in the literature include meta-norms and second-order norms, which describe expected sanctioning behavior and expected responses to sanctions (e.g. Chaudhuri, 2011; Cinyabuguma, Page, & Putterman, 2006; Irwin & Horne, 2013).

It is easy to think of oppositional control as an inherently antisocial act, expressed for example as antisocial punishment or revenge. In the context of experimental research on social dilemmas, this interpretation is often justified. These experiments tend to have clearly defined collectively optimal behavior (cooperation in a prisoner's dilemma, contributions in a public good game), and sanctioning systems intended to enforce this collectively optimal behavior. Exercising oppositional control in this context generally goes against the interests of the group as a whole. Indeed, in the present study, we treat vulnerability to oppositional control as a negative. Sanctioning systems vulnerable to oppositional control may be less effective at enforcing high rates of cooperation.

However, it is important to note that oppositional control is not universally opposed to the collective interest. For example, sanctioning systems can be abused to enforce behavior which is beneficial to those distributing sanctions but not to the group as a whole. This applies to situations of overprovision of a public good, where the cost of producing a public good exceeds its value (Heckathorn, 1993). This also

applies to situations where a group consists of several subgroups with heterogeneous interests, where oppositional control may prevent one subgroup from forcing another to act against its interests. For example, environmentally conscious behavior enforced by a tax on produce which is not grown in an environmentally friendly way contributes to the solution of an important collective problem, but the added costs may be too much to bear for those with lower incomes. The ability to exercise oppositional control may prevent these individuals from having to comply with legislation which would substantially decrease their standard of living.

Legitimacy of punishment

If oppositional control is used to prevent or discourage compliant control, then an important factor in the decision to exercise oppositional control may be the extent to which compliant control is considered legitimate. The legitimacy of compliant control influences how likely recipients of sanctions are to change their behavior (Baldassarri & Grossman, 2011), and likely also influences the extent to which group members decide to exercise oppositional control. The legitimacy of punishment depends, among other things, on the extent to which compliant control is collectivized (Strimling & Eriksson, 2014) and on the basis from which punishers derive their authority (Baldassarri & Grossman, 2011; Grossman & Baldassarri, 2012).

Institutional punishment systems are likely to be considered more legitimate than peer punishments, because compliant control is more collectivized and the punishers' authority derives from the community. Sanctions distributed by institutional punishment systems are of a fundamentally different nature than those distributed by peers. Institutional punishers take care to ensure that sanctions are distributed in the name of the community as a whole (c.f. phrases such as 'the people versus' or 'the crown versus' in criminal proceedings). Acceptable and unacceptable behavior have been defined in advance, together with the corresponding sanctions, by the community as a whole. Because institutional punishment is (at least in its ideal cases) fundamentally impersonal, sanctions cannot be driven by the punisher's subjective desire to inflict punishment on a particular victim. Personal grudges, therefore, play a much smaller role in collective punishment than in peer punishment. Hence, from the perspective of the punished, collective punishments are more likely to be perceived as communal acts of moral righteousness rather than acts of personal vengeance (Baldassarri & Grossman, 2011). Generally, victims will, therefore, more readily accept institutional punishments as justified, leading to more compliance and less oppositional control.

As a first consequence of the higher legitimacy of institutional punishment, we expect that individuals who receive punishment from an institution are more likely to adjust their behavior than those who receive punishment from peers. That is, punished participants will become more likely to cooperate in subsequent rounds and this effect will be stronger under an institutional punishment system.

Hypothesis 1: The probability that a participant cooperates increases more after receiving collective punishment than after receiving peer punishment

Oppositional control

Based on experimental studies, peer punishment has been established as an effective tool to increase cooperation in social dilemmas (Balliet et al., 2011; Chaudhuri, 2011; Fehr & Gächter, 2002; Flache et al., 2017). However, peer punishment systems restrict neither who is allowed to distribute punishments nor at whom these punishments are allowed to be directed. As a result, peer punishment systems allow both antisocial punishment (punishment directed at those who exhibit collectively beneficial behavior (Chaudhuri, 2011; Guala, 2012; Herrmann et al., 2008; Irwin & Horne, 2013)) and retaliation (punishment directed at those who exercise compliant control (Chaudhuri, 2011; Guala, 2012; Nikiforakis, 2008; Nikiforakis et al., 2012)). Retaliation is a form of oppositional control which has the potential to drastically decrease the effectiveness of peer punishment as a sanctioning system (Nikiforakis, 2008; but see Chapter 3 of this thesis for a counter-example). Retaliation can discourage punishment, and persistent punishers may be drawn into long and costly feuds (Nikiforakis & Engelmann, 2011; Nikiforakis et al., 2012).

Institutional punishment systems also provide opportunities for oppositional control, but these opportunities are more limited and of a different nature. For example, institutional punishment systems are less personal than peer punishment systems, limiting opportunities for retaliation. Fines may even be distributed by automated systems (e.g. speed cameras), not exposing any individual within the institution to retaliation. Sometimes retaliation is possible, such as when an offender receiving a citation from a police officer responds by verbally or physically assaulting the officer. Even then, this act of retaliation likely does little to discourage future punishment from the institution as a whole.

Nonetheless, even institutional punishment systems are to some extent vulnerable to oppositional control. Oppositional control in the case of institutional punishment involves decreasing the ability of the institution to effectively exert

control. For instance, dissatisfaction with the police and concern that police forces have too much leeway in their sanctioning behavior may result in decreased police budgets and more limitations on the use of force. Such limitations can be achieved through legislation or litigation, under the pressure of public opinion. Notably, this type of oppositional control requires dissatisfaction with the institution as a whole rather than with any one of its agents. Even when a public response is provoked by the behavior of individual police officers, this does not interfere with the institution's ability to exert control unless the institution itself is seen as complicit.

Comparing peer punishment systems and institutional punishment systems

Based on our assessment of opportunities for oppositional control in both types of sanctioning systems, we expect that both peer punishment systems and institutional punishment systems perform better when oppositional control is precluded. In the context of the prisoner's dilemma game, we investigate in this study, the performance of a sanctioning system is expressed through the cooperation rate attained. We, therefore, expect the following:

Hypothesis 2: Cooperation rates will be higher in a peer punishment system without oppositional control than in a peer punishment system with oppositional control

Hypothesis 3: Cooperation rates will be higher in a collective punishment system without oppositional control than in a collective punishment system with oppositional control

However, we do not expect that both sanctioning systems are affected equally. Major contributing factors in this difference are likely the higher legitimacy of institutional punishment and its less personal nature. Peer punishment systems generally allow more opportunities for oppositional control than institutional punishment systems. And, even assuming equal opportunities for oppositional control, we expect less oppositional control to be exercised in institutional punishment systems because institutional punishments are likely to be seen as more legitimate. Both factors will discourage oppositional control and therefore limit the negative impact of oppositional control on cooperation rates.

Hypothesis 4: The decrease in cooperation from the possibility of oppositional control will be greater under a peer punishment system than under a collective punishment system

EXPERIMENT

The basis of our experiment is formed by a 5-person Prisoners Dilemma with punishment. Participants play 10 rounds, and the composition of the group is constant throughout the 10 rounds. However, at the start of each round participants are relabeled. Participants, therefore, know that they are playing with the same four others for 10 rounds, but cannot develop personal reputations based on decisions in previous rounds. There are five experimental treatments, which will be discussed in detail below. Table 4.1 gives an overview of the important features of each treatment. Below, we describe the experiment and treatments in more detail. During the experiment, participants earn points, which are converted to Euros at the end of the experiment.

Table 4.1. Overview of treatments

Treatment	*Type of compliant control*	*Type of oppositional control*
Peer punishment without oppositional control	Peer punishment	None
Peer punishment with retaliation	Peer punishment	Retaliation
Institutional punishment without oppositional control	Institutional punishment through punishment pool	None
Institutional punishment with retaliation	Institutional punishment through punishment pool	Retaliation
Institutional punishment with removal	Institutional punishment through punishment pool	Removing punishments from the pool

Cooperate or defect

The first decision participants make is whether to cooperate or defect. This decision is identical across all five treatments. Choosing to cooperate costs 20 points and results in a total yield for the group of 40 points (i.e. 8 points per group member). Thus, choosing to cooperate is costly to the individual but beneficial for the group as a whole. All participants make this decision simultaneously and in ignorance of what the others are choosing.

Compliant control

In all five treatments, the decision to cooperate or defect is followed by the opportunity to exercise compliant control. In all treatments, this stage starts by displaying to each participant the decisions made by each member of their group. Participants can tell how many group members chose to cooperate, how many chose to defect, and which decision each individual group member made.

In the peer punishment treatments, participants have the opportunity to punish their group members. For each group member who chose to defect, participants have to select whether or not they want to punish this person. Restricting punishment to group members who defected precludes antisocial punishment and thus ensures that punishments distributed by peers are directed at the same persons (defectors) who would be targeted by the institutional punishment system. Punishment is costly: participants pay a cost of 5 points for each punishment they distribute. A player who receives a punishment loses 5 points for each punishment they receive.

In the institutional punishment treatments, participants have the opportunity to invest in an automated punishment institution. Rather than deciding whether or not to punish each individual defecting group member, participants can contribute a number of punishments to a punishment pool. The maximum number of punishments a participant can contribute is equal to the number of group members who defected. Contributing a punishment to the pool is costly: participants pay a cost of 5 points for each punishment they distribute. The automated punishment institution then distributes the punishments contributed to the pool randomly among defectors, without replacement. That is, each defector can only be punished once from the punishments contributed to the pool by a particular punisher. Recalling that under the peer punishment system every player can punish each defecting other exactly once, this restriction renders institutional punishment and peer punishment equivalent in this respect. As a result, if you contribute a number of punishments to the pool equal to the number of defectors, you know that the punishments you contributed will be distributed such that each defector receives exactly one of your punishments. A player who receives a punishment also loses 5 points.

Note that the costs, effectiveness, and constraints on the distribution of punishment are identical between the peer punishment system and the institutional punishment treatment. The only difference between the systems is whether the punishment is given directly from peer to peer, or distributed by a centralized institution.

Oppositional control

In three treatments, one involving peer punishment and two involving institutional punishment, there are opportunities for oppositional control. The implementation of oppositional control differs across these three treatments.

Under the peer punishment system, oppositional control involves retaliation against punishers. In this treatment with peer punishment and oppositional control, after the compliant control stage, all participants are informed whether any group members punished them and if so, which group members these were. They can then choose to retaliate by distributing counter-punishments. Counter-punishments can only be directed at group members who punished you. For each punisher, punished participants have to select whether or not they want to retaliate against this punisher. The cost and effectiveness of counter-punishment are identical to that of punishment. A player who receives a counter-punishment loses 5 points.

Under the institutional punishment system, oppositional control is implemented in two different ways. First, in the treatment with institutional punishment and retaliation, punished players can retaliate by distributing counter-punishments as in the peer punishment system. All participants are informed whether they received any punishments from the institution, and are also informed which participant contributed *that particular punishment* to the pool. They are then given the opportunity to retaliate against this person. Counter-punishment is otherwise as described for the peer punishment system. The information received about other participants and the type of oppositional control possible are thus identical to counter-punishment under the peer punishment system. The only difference is that under the institutional punishment system, the participant who contributed a given punishment to the pool did not target any particular other (since the automated punishment system randomly selects the target) but instead expressed disapproval of non-cooperators in general.

Second, in the treatment with institutional punishment and removal, oppositional control consists of removing punishments from the pool, before they are inflicted. That is, by paying a removal fee of 5 points, participants can prevent one punishment from occurring. Punishments to prevent are selected randomly after all participants have decided how many punishments to remove from the pool. Punished participants thus cannot reliably prevent punishments of themselves, and in any case, it would not be economical to do so since the removal fee is equal to the fine they would incur when punished. Note, however, that the ability to remove punishments

from the pool is not restricted to participants who have themselves been punished. Anyone can remove punishments from the pool.

This type of oppositional control differs in several ways from counter-punishment. First, removal of punishments from the pool happens before punishments have been dispensed. This form of oppositional control thus cannot be motivated by anger at being punished but can be motivated by a fear of being punished or by disapproval of punishment in general. Second, removing punishments from the pool does not harm the person who contributed this punishment (except in the sense that the punishment they contributed is ineffective) while counter-punishment involves imposing additional costs on punishers. Third, this form of oppositional control also allows oppositional control to be exercised by individuals who were not themselves subject to punishment. This is thus 'true' oppositional control, in the sense that all possible strategies including compliant opposition (Heckathorn, 1993) are possible. As a result, this method of oppositional control may be more realistic than retaliation in institutional punishment systems, but may also be less effective at preventing oppositional control than direct counter-punishment.

With this additional treatment, our experiment allows us to both investigate the robustness of our results to the exact implementation of oppositional control and explore different theoretical mechanisms which explain why institutional punishment systems may be less vulnerable to oppositional control than peer punishment systems are. If institutional punishment is less vulnerable to oppositional control only because it makes oppositional control less effective, then we should observe a difference between institutional punishment and peer punishment only for the institutional punishment treatment where opposition means removal of punishments. If, on the other hand, the difference is due to differences in legitimacy, then we should observe a difference between institutional punishment and peer punishment also for the institutional punishment treatment where opposition means retaliation. If both factors are important, we should observe a difference compared to peer punishment in either case, and the difference should be larger when we limit opportunities for oppositional control.

Procedure

The experiments were conducted at the Sociological Laboratory of the University of Groningen (http://www.soclab.nl). The Sociological Laboratory has a subject pool consisting mainly of students at the University of Groningen. These students come from a variety of disciplines including sociology, economics, law, biology, physics, etc.

Within the subject pool, psychology students and sociology students are overrepresented, compared to the population of students at the University. A small number of non-students are also registered with the Sociological Laboratory and participate in experiments. The rules of the Sociological Laboratory guarantee subjects that they will not be deceived in the experiment, and that they will be paid for their efforts.

Experiments took place in computer rooms prepared in such a way that, once they were seated, participants could not see the screens on which the other participants were playing. The experiments were programmed using zTree software for economic experiments (Fischbacher, 2007). The experiment started with an introduction by the experimenter, explaining the rules of conduct within the lab and asking the participants to start reading the instructions. Before the main experiment, which for each participant consisted of one of the treatments described above, participants filled out a non-incentivized measure of Social Value Orientation (the 9-item triple dominance measure; Au & Kwong, 2004; Van Lange, Otten, De Bruin, & Joireman, 1997) and a more general questionnaire on norms regarding cooperative behavior.

Instructions were provided on participants' screens, and were repeated before each round of the experiment. During the experiment, subjects were always allowed to take notes. This ensured that in the treatments where participants might want to remember information across rounds (e.g. details of the instructions, outcomes from previous rounds) they would not be required to memorize this information.

During the experiment, participants earned points depending on their decisions and those of their group members. At the end of the experiment, these points were converted to Euros at a fixed rate, such that on average participants earned between €2 and €3 from the experiment, on top of their €5 show-up fee.

Data

Data were collected from a total of 160 participants across 12 sessions. One session had to be terminated after three rounds due to a software issue. Data from this session have been excluded, resulting in a final total of 145 participants. All in all, we collected data on 1450 cooperation decisions, 1198 punishment opportunities (708 opportunities for punishment in the peer punishment treatments, 490 opportunities to contribute punishments to the pool in the institutional punishment treatments), and 131 opportunities for oppositional control (100 retaliation decisions and 31 removal decisions). These opportunities will form the units of analysis. Each case is

nested in an individual (the participant who took the decision) and a group (the group this participant was part of at the time of the decision). This multilevel structure will be taken into account in all statistical analyses performed in the following section. We estimate multilevel models using the R package lme4 (Bates et al., 2015). In each analysis, we include uncorrelated random intercepts for participants and groups, in addition to the level-1 error.

RESULTS

Social Value Orientation

We include the 9-item Triple Dominance Measure of Social Value Orientation in our analyses as a control variable. Not only are individuals' Social Value Orientations predictive of their behavior in social dilemmas (Bogaert et al., 2008; Pletzer et al., 2018), the distribution of Social Value Orientations in the group involved in a social dilemma may have an additional impact on collective outcomes (Chapter 3). Table 4.2 shows the distribution of Social Value Orientations in our sample. The majority of our participants has a Prosocial orientation, which is consistent with other studies from our laboratory (e.g. Dijkstra & Bakker, 2017; Chapter 3 of this dissertation) and with studies using other Dutch samples (Van Lange et al., 1997, 2011) or samples from other countries (Au & Kwong, 2004).

Table 4.2. Social Value Orientations

	Frequency	*Percentage*
Prosocial	80	55.17 %
Individualistic	45	31.03 %
Competitive	4	2.76 %
Mixed / Unclassified	16	11.03 %
Total	145	100 %

The effectiveness of peer punishment and institutional punishment

Hypothesis 1 predicts that the effect of received punishment on the probability that a participant cooperates is greater for institutional punishment than for peer punishment. We test this hypothesis using a multilevel logistic regression model. The 1305 cooperation decisions our participants made in the second through last rounds of all treatments are the units of analysis in this model. We exclude cooperation decisions from the first round because we are trying to determine how participants

change their behavior in response to punishment. In the first round, participants cannot have received any punishment yet. The dependent variable in this model is the decision to cooperate (1) or defect (0). There are three main independent variables of interest in this model. First, whether the sanctioning system was based on peer punishment or institutional punishment. Second, how many punishments the participant received in the previous round. Third, an interaction term between the sanctioning system and the received punishment which will be used to test Hypothesis 1. We control for the cooperation decision the participant made in the previous round, as we not only expect this to be relatively stable across rounds but also to be strongly correlated with the punishment this participant received in the previous round. Additionally, we control for the Social Value Orientation of the participant through three dummy variables representing the three measured orientations (Prosocial, Individualistic and Competitive). The remaining unclassified participants form the reference category.

Table 4.3 shows the results of this model, as well as the results from a model including only the control variables. The most important predictor of a participant's cooperation decision is their decision from the previous round. Those who cooperated before are more likely to cooperate again (b = 1.469, $p < 0.001$ one-sided). Under the peer punishment system, receiving punishment does not seem to have any additional effect on the probability that a participant cooperates (b = -0.001, p = 0.499 one-sided). The interaction between the sanctioning system and the received punishment from the previous round shows that punishment has a larger effect on cooperation under the institutional punishment system (b = 0.492, p = 0.034 one-sided), consistent with Hypothesis 1. Running the model again with the institutional punishment system as reference category shows that under institutional punishment participants do cooperate significantly more often when they received more punishment in the round before (b = 0.491, p = 0.010).

Result 1: Contributions are increased after receiving institutional punishment and not after receiving peer punishment.

In interpreting this result, we should note that for this analysis we conditioned on posttreatment variables (Montgomery, Nyhan, & Torres, 2018) by controlling for participants' decision to cooperate or defect. Participants who defect under a peer punishment system may be systematically different from participants who defect under an institutional punishment system and may respond differently to punishment

because of those differences rather than because of the difference in punishment system.

Table 4.3. Estimates of multilevel logistic regression models for cooperation

	Model 1	*Model 2*
Treatment variables		
Institutional punishment		-0.090 (0.588)
Number of punishments received [a]		-0.001 (0.212)
Institutional punishment x Punishments received		0.492 (0.270)*
Control variables		
Cooperation decision [a]	1.304(0.168)***	1.469 (0.202)***
SVO (Compared to Unclassified)		
-- Prosocial	0.328 (0.255)	0.330 (0.256)
-- Individualistic	-0.505 (0.275)*	-0.526 (0.276)*
-- Competitive	0.649 (0.671)	0.649 (0.675)
Constant	0.292 (0.391)	0.187 (0.536)
Random effects (SD)		
Subject	0.000	0.000
Group	1.477	1.464
N	1305	1305
Deviance	1180.2	1174.6

*Note. [a] In the previous round; One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

Oppositional control in peer punishment institutions

Hypothesis 2 predicts less cooperation in peer punishment systems which allow oppositional control through retaliation than in peer punishment systems without oppositional control. Descriptively, our results appear to be consistent with this hypothesis. In the peer punishment treatment without oppositional control, 85% of cooperation decisions resulted in cooperation (N = 300). In the peer punishment treatment with oppositional control, in the form of retaliation, only 56% of cooperation decisions resulted in cooperation (N = 300).

To test this hypothesis, we estimate a multilevel logistic regression model. The 600 cooperation decisions our participants made in the two peer punishment treatments are the units of analysis in this model. The dependent variable in this model is the decision to cooperate (1) or defect (0). The main independent variable in this model is the treatment, specifically whether oppositional control through

retaliation was possible. Additionally, we control for the Social Value Orientation of the participant through three dummy variables representing the three measured orientations (Prosocial, Individualistic and Competitive). The remaining unclassified participants form the reference category.

Table 4.4 shows the results of this model, as well as a model including only the control variables. We find that participants cooperated significantly less frequently under a peer punishment system with retaliation than under a peer punishment system without retaliation (b = -2.457, p = 0.003 one-sided).

Result 2: Cooperation is significantly less likely under a peer punishment system which allows oppositional control, through retaliation, than under a peer punishment system which does not allow oppositional control.

Table 4.4. Estimates of multilevel logistic regression models for cooperation in peer punishment systems

	Model 1	*Model 2*
Treatment variables		
Oppositional control		-2.457 (0.883)**
Control variables		
SVO (Compared to Unclassified)		
-- Prosocial	-0.002 (0.471)	0.021 (0.467)
-- Individualistic	-1.186 (0.507)**	-1.172 (0.505)*
-- Competitive	0.498 (0.793)	0.510 (0.794)
Constant	1.880 (0.713)	3.096 (0.782)
Random effects (SD)		
Subject	0.000	0.000
Group	1.816	1.420
N	600	600
Deviance	555.3	540.6*

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

Oppositional control in institutional punishment systems

Hypothesis 3 predicts less cooperation in institutional punishment systems which allow oppositional control, through either removal or retaliation, than in institutional punishment systems without oppositional control. Our results are not consistent with this hypothesis. In the institutional punishment treatment without oppositional

control, 66% of cooperation decisions resulted in cooperation (N = 350). In the institutional punishment treatments with oppositional control 75.8% of cooperation decisions resulted in cooperation (N = 500). The cooperation rate is highest when oppositional control is operationalized as retaliation against those who contributed to the punishment institution (78.8%, N = 250).

To formally test Hypothesis 3, we again estimate a multilevel logistic regression model. The 850 cooperation decisions our participants made in the three collective punishment treatments are the units of analysis in this model. The dependent variable in this model is the decision to cooperate (1) or defect (0). The main independent variable in this model is the treatment, specifically whether oppositional control was possible, without distinguishing between retaliation and the removal of punishments. Additionally, we control for the Social Value Orientation of the participant through three dummy variables representing the three measured orientations (Prosocial, Individualistic and Competitive). The remaining unclassified participants form the reference category.

Table 4.5 shows the results of this model, as well as a model including only the control variables. We find that participants did not cooperate significantly more or less frequently under an institutional punishment system with oppositional control than under a peer punishment system without oppositional control (b = 0.834, p = 0.316 one-sided). Testing differences between the treatment without oppositional control and each of the two treatments with oppositional control separately also does not result in any significant differences.

Result 3: Cooperation is not significantly more or less likely under collective punishment systems with oppositional control than under collective punishment systems without oppositional control.

Table 4.5. Estimates of multilevel logistic regression models for cooperation in peer punishment systems

	Model 1	*Model 2*
Treatment variables		
Oppositional control		0.834 (0.832)
Control variables		
SVO (Compared to Unclassified)		
-- Prosocial	0.535 (0.345)	0.534 (0.345)
-- Individualistic	-0.479 (0.366)	-0.489 (0.366)
-- Competitive	0.245 (1.303)	0.207 (1.305)
Constant	1.365 (0.518)	0.873 (0.690)
Random effects (SD)		
Subject	0.494	0.495
Group	1.658	1.596
N	850	600
Deviance	818.2	817.2

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

Comparing the effects of oppositional control in peer punishment systems and collective punishment systems

Hypothesis 4 predicts that the negative difference between cooperation rates in sanctioning systems with and without oppositional control should be larger in peer punishment systems than in institutional punishment systems. As we have already seen, cooperation rates are clearly lower in the peer punishment system with oppositional control than in the peer punishment system without. We have also seen that there is no significant difference between treatments with and without oppositional control when it comes to institutional punishment systems.

We formally test Hypothesis 4 using a multilevel logistic regression model. The 1450 cooperation decisions our participants made across all treatments are the units of analysis in this model. The dependent variable in this model is the decision to cooperate (1) or defect (0). There are three main variables of interest in this model. First, whether the sanctioning system was based on peer punishment (reference) or institutional punishment. Second, whether oppositional control was or was not (reference) possible. Third, an interaction term between the sanctioning system and the possibility of oppositional control which will be used to test Hypothesis 4.

Additionally, we control for the Social Value Orientation of the participant through three dummy variables representing the three measured orientations (Prosocial, Individualistic and Competitive). The remaining unclassified participants form the reference category.

In this analysis, we do not distinguish between the two types of oppositional control which existed under the institutional punishment system in this experiment. After all, in the peer punishment system only retaliation is possible, and in the institutional punishment system there is no significant difference in cooperation rates between the treatments with different types of oppositional control. [4]

Table 4.6 shows the results of this model, as well as the results from a model including only the control variables. We find that participants cooperated significantly less frequently under the institutional punishment system without oppositional control than under the peer punishment system without oppositional control (b = -1.795, p = 0.024 one-sided). As we saw previously, participants also cooperated significantly less frequently under the peer punishment system with oppositional control than under the peer punishment system without oppositional control (b = - 2.500, p = 0.004 one-sided). Consistent with Hypothesis 4, there is a significantly positive interaction between the presence of oppositional control and the type of sanctioning system (b = 3.324, p = 0.003 one-sided). The value of the interaction coefficient indicates that in institutional punishment systems the cooperation rate is, in fact, higher when oppositional control is possible than when it is not. However, as we know from our previous analyses, this difference is relatively small and is not statistically significant. When we look only at the 800 cases from treatments with oppositional control, cooperation rates were higher under an institutional punishment system than under a peer punishment system (b = 1.516, p = 0.024 one-sided).

Result 4: There is less cooperation in treatments with institutional punishment than in a peer punishment system without oppositional control. The possibility of oppositional control negatively affects cooperation rates under a peer punishment system but not under an institutional punishment system.

[4] If instead we omit the treatment with institutional punishment in which oppositional control is operationalized as removal of punishments, ensuring minimal differences between the treatments being compared, none of the coefficients are substantially changed. There are no differences in either direction or significance of coefficients for the treatment variables.

Table 4.6. Estimates of multilevel logistic regression models for cooperation in peer punishment systems

	Model 1	*Model 2*
Treatment variables		
Institutional punishment		-1.795 (0.904)*
Oppositional control		-2.501 (0.930)**
Institutional punishment x Oppositional control		3.324 (1.219)**
Control variables		
SVO (Compared to Unclassified)		
-- Prosocial	0.385 (0.268)	0.404 (0.269)
-- Individualistic	-0.670 (0.286)**	-0.669 (0.286)**
-- Competitive	0.732 (0.692)	0.739 (0.701)
Constant	1.494 (0.416)	2.757 (0.722)
Random effects (SD)		
Subject	0.380	0.383
Group	1.757	1.505
N	1450	1450
Deviance	1368.8	1361.4

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

DISCUSSION

It has been proposed that one of the reasons why peer punishment systems are not a common part of social life is that peer punishment systems are particularly vulnerable to oppositional control (Chaudhuri, 2011; Guala, 2012; Nikiforakis, 2008). While the vulnerability of peer punishment to oppositional control has been established (Nikiforakis, 2008), the relative invulnerability of institutional punishment has thus far been assumed. In the present study, we compared a peer punishment system and an institutional punishment system, with and without opportunities for oppositional control. Our results provide empirical support for the expectation that institutional punishment systems are less vulnerable to oppositional control than peer punishment systems.

In peer punishment systems, opportunities for oppositional control are implemented as opportunities for retaliation (Chaudhuri, 2011; Nikiforakis, 2008; Nikiforakis et al., 2012). In peer punishment systems, the burden of distributing punishment falls on each group member individually. There is no coordination of

punishment, and punishments are given directly from one person to another. Under these conditions, retaliation may be frequent and effective. Punishment under a peer punishment system is very personal. Retaliation may be frequent because recipients of punishment may feel unfairly treated, may feel that their peers should not have the authority to distribute punishments, and may feel personally attacked. At the same time, retaliation may be effective. Punishers may be afraid of retaliation and given that punishment is not coordinated may be tempted to hope that another group member will act as punisher.

In institutional punishment systems, oppositional control may be less frequent, less impactful, and fundamentally of a different nature. Oppositional control under an institutional punishment system is more likely to consist of the prevention of punishment than of retaliation against punishers. This is both because the punishment is often distributed by impersonal entities (such as tickets from automated speed cameras) and because there are often strong legal and social consequences for retaliating against representatives of an institution (such as harsh punishment for assaulting a police officer). At the same time, regardless of how oppositional control may take place, punishment received under an institutional punishment system is likely seen as more legitimate and, as a result, recipients of punishment may be less inclined to exercise oppositional control.

Conclusions

We examined differences between peer punishment systems and institutional punishment systems experimentally, in a 5-person prisoner's dilemma game. We expected that the possibility of oppositional control would negatively affect cooperation rates both in peer punishment systems and in institutional punishment systems. We found that participants were indeed less likely to cooperate under a peer punishment system with oppositional control than in one without, consistent with *Hypothesis 2*. However, in contrast to *Hypothesis 3*, the possibility of oppositional control did not seem to affect cooperation in an institutional punishment system. While this second result is somewhat unexpected, it does align with our expectation that the negative impact of oppositional control is greater in peer punishment systems than in institutional punishment systems (*Hypothesis 4*). The expected difference in the impact of oppositional control was confirmed by a separate analysis. Notably, a peer punishment system without opportunities for oppositional control displayed higher rates of cooperation than an institutional punishment system without oppositional control. This occurred in spite of our observation that institutional

punishment, when received, has a stronger positive impact on the punished participant's future contributions than peer punishment has. However, when oppositional control is possible an institutional punishment system results in higher cooperation rates than a peer punishment system.

We also expected that the effectiveness of punishment would be higher when this punishment is distributed by an institution rather than by an individual. Our results show that consistent with *Hypothesis 1*, participants who received punishment under an institutional punishment system were more likely to cooperate in the subsequent round while participants who received punishment under a peer punishment system were not.

It appears that under ideal conditions, where groups can exercise compliant control and there is no possibility of oppositional control, peer punishment can be more effective than institutional punishment. This may well be because in peer punishment systems each group member has an individual motivation to distribute punishment. While supporting a punishment institution may be something that people find valuable and worthwhile, distributing punishment personally comes with a degree of emotional satisfaction that an institution cannot match. The provision of an effective sanctioning system is itself a public good, as it requires individual contributions to enable the distribution of sanctions, while the benefit of increased cooperation can be enjoyed by everyone (Yamagishi, 1986). The emotional satisfaction from personally punishing a non-contributor may be a selective incentive (Olson, 1965) which encourages contributions to this second-order public good, making second-order free riding (Heckathorn, 1989; Oliver, 1980) more prevalent in institutional punishment systems than in peer punishment systems.

However, there are distinct advantages to an institutional punishment system when group members *can* exercise oppositional control. No matter whether the oppositional control consists of removing support for the distribution of punishment or of retaliating against individuals who supported the institution, cooperation rates are unaffected. At least in part, this may be because institutional punishment is considered more legitimate than peer punishment. If institutional punishment is considered more legitimate, then group members may be less likely to make use of opportunities for oppositional control as well as being more responsive to the punishment they receive. Two outcomes of the present study appear to support the importance of legitimacy in explaining differences between the two sanctioning systems. For one, punishment is more effective at enforcing contributions under an institutional punishment system than under a peer punishment system. That is,

participants become more likely to contribute after receiving punishment from an institution, but not after receiving punishment from one of their peers. Second, there was no significant difference in cooperation rates between the two institutional punishment treatments with different types of oppositional control. This may indicate that institutional punishment is less vulnerable to oppositional control not because oppositional control is often less effective in institutional punishment systems, but rather because participants were less keen to exercise oppositional control.

Discussion, limitations, and directions for future research

While it appears from the present study that institutional punishment systems may be preferable to peer punishment systems, because they avoid vulnerability to oppositional control which is likely to severely hamper peer punishment in most situations, there are several other factors to consider.

First, there are many costs and benefits which differ between peer punishment systems and institutional punishment systems, which we have not considered in this study. For example, peer punishment systems are far easier to implement than institutional punishment systems. In fact, peer punishment systems can be said to implement themselves by default whenever there is no clear authority governing a particular situation. Establishing an institutional punishment system requires legitimate authority, consensus on which rules should be enforced and what punishment should be distributed, and the recruitment of enforcement agents. Institutional punishment systems are often relatively inflexible and may find it difficult to effectively monitor all relevant actions (Erikson & Parent, 2007). Institutional punishment systems also require consistent investment, which may be difficult to adjust in response to the number of offenders (Traulsen et al., 2012). There are significant costs and benefits to both types of sanctioning systems, and their vulnerability to oppositional control is just one of many to consider.

Second, resilience to oppositional control cannot always be considered a positive property of a sanctioning system. When we can be reasonably sure that sanctions will be directed at individuals acting against the interests of everyone else in a social group, such as in the stylized situation of the n-person prisoner's dilemma, this seems reasonable. However, in other situations, a lack of oppositional control may lead to overprovision of public goods (Heckathorn, 1993) or oppression of one subgroup by another. In extreme situations, institutions designed to distribute sanctions on which there is consensus within the group may be taken over by outside actors or may decide not to comply with the desires of the group as a whole. In other

words, peer punishment systems appear to be more unstable than institutional punishment systems, and this may be good or bad depending on the circumstances.

Third, from the present study, it is still difficult to draw conclusions on why exactly peer punishment is more strongly affected by the possibility of oppositional control than institutional punishment is. It may be reasonable to expect that this is because individuals fear retaliation more in the peer punishment system than they fear either form of oppositional control in the institutional punishment system. Our results are consistent with the idea that institutional punishment is considered more legitimate than peer punishment, which would lead to the expectations that retaliation against institutional punishment is uncommon. However, the number of instances in which any form of opposition took place was too small to analyze with sufficient statistical power. In addition, we did not measure the perceived legitimacy of the two types of sanctioning systems.

Fourth, in the interest of ensuring maximum comparability between all treatments, we made a number of decisions which may seem particularly unrealistic. For example, we assumed that distributing punishment is as costly to the punisher as to the person being punished. We also assumed that the effort of preventing a punishment was as costly as being punished. Our results may have been influenced by these decisions. For example, had punishment been more cost-effective it may have been more difficult to discourage potential punishers through oppositional control. If punishment was sufficiently efficient even after taking into account the effects of oppositional control, cooperation rates may have been equal across all treatments. This is a consideration when attempting to generalize to from this study.

We want to point towards three potential directions for future research on this topic. First, keeping in mind that we cannot conclusively say why cooperation rates differ across the treatments of our experiment, we hope that future research can explore this further. One might consider testing explicitly some of the assumptions we have made in the present study, such as the idea that higher legitimacy can explain different responses to peer punishment and institutional punishment. This was not the focus of the present study, as we studied observed behavior rather than the motivations behind this behavior. Second, keeping in mind the decisions we made in the design of our experiment, it may be valuable to replicate the present study both exactly and with changes to some aspects of our design about which we did not explicitly theorize. If the observed differences persist, we can be more confident that our results are not highly sensitive to the details of the experimental design. Third, keeping in mind that many other factors affect the costs and benefits of peer

punishment systems and institutional punishment systems, we hope that future research will broaden the scope of the comparison of these sanctioning systems. We have pointed to some examples already, such as groups with heterogeneity and cases of overprovision. In addition, future research may attempt to bring in more of the costs and benefits of the two sanctioning systems, perhaps endogenously establishing the sanctioning rules carried out by the institution.