Marcin, Isabel; Robalo, Pedro; Tausch, Franziska

Working Paper

Institutional endogeneity and third-party punishment in social dilemmas

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2016/6

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Marcin, Isabel; Robalo, Pedro; Tausch, Franziska (2016) : Institutional endogeneity and third-party punishment in social dilemmas, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2016/6, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/144910

Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2016/6

Institutional Endogeneity and Third-party Punishment in Social Dilemmas

Isabel Marcin
Pedro Robalo
Franziska Tausch

MAX PLANCK SOCIETY


Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2016/6

Institutional Endogeneity and Third-party Punishment in Social Dilemmas

Isabel Marcin / Pedro Robalo / Franziska Tausch

April 2016

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Institutional Endogeneity and Third-party Punishment in Social Dilemmas

Isabel Marcin, Pedro Robalo and Franziska Tausch *

April 18, 2016

Abstract

This paper studies experimentally how the endogeneity of sanctioning institutions affects the severity of punishment in social dilemmas. We allow individuals to vote on the introduction of third-party-administered sanctions, and compare situations in which the adoption of this institution is endogenously decided via majority voting to situations in which it is exogenously imposed by the experimenter. Our experimental design addresses the self-selection and signaling effects that arise when subjects can vote on the institutional setting. We find that punishment is significantly higher when the sanctioning institution is exogenous, which can be explained by a difference in the effectiveness of punishment. Subjects respond to punishment more strongly when the sanctioning institution is endogenously chosen. As a result, a given cooperation level can be reached through milder punishment when third-party sanctions are endogenous. However, overall efficiency does not differ across the two settings as the stricter punishment implemented in the exogenous one sustains high cooperation as subjects interact repeatedly.

Keywords: Endogeneity · Third-party punishment · Voting · Institutions · Social dilemma · Public good

JEL Classification: C92 · D02 · D72 · H41

*Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str.10, 53113 Bonn, Germany. E-mail: Marcin: marcin@coll.mpg.de, Robalo: robalo@coll.mpg.de, Tausch (corresponding author): tausch@coll.mpg.de.

1 Introduction

Institutions are a crucial determinant of social interaction outcomes. In situations in which self-interest clashes with collective interest - so-called social dilemmas - human societies have developed a number of institutional arrangements that mitigate the inherent free-rider problem. The third-party enforcement of social norms is one such arrangement that has shown to be successful in enhancing cooperation (Charness et al., 2008; Lergetporer et al., 2014). In this paper we study the extent to which the process that generates a third-party punishment institution influences its severity and how the affected individuals respond to it. In particular, we compare settings in which groups select third-party sanctions through majority voting with settings in which sanctions are exogenously put in place. This allows us to investigate whether third-parties who have been introduced through majority voting enforce cooperation norms differently than those who have been exogenously appointed.

The available empirical evidence shows that endogenous punishment institutions lead to more cooperative behavior than their exogenous counterparts. This phenomenon is typically referred to as the 'endogeneity effect'. In particular, institutions tend to be more effective in increasing cooperation when individuals can determine their implementation through majority voting (e.g. Tyran and Feld, 2006). In principle, this effect could be due to the self-selection of cooperative individuals into their preferred institution, to the cooperation signal inherent to the voting outcome, or to the concession of participation rights to individuals. Several experimental papers show that endogeneity increases cooperation even after controlling for the selection and signaling channels (e.g. Dal Bó et al., 2010; Kamei et al., 2015). This regularity is referred to as the 'endogeneity premium': allowing groups to adopt sanction or reward schemes drives an increase in cooperative behavior.

While the existing studies on centralized institutions involve punishment levels that are predetermined and executed by an automatic mechanism, in reality punishment is the product of human judgment and is frequently administered by third parties. In most cases punishment is carried out under more or less established sets of rules (social norms, legal rules, etc.), but leeway is granted to the authority by whom it is administered. One prominent example is the judicial system: judges are bound by law but decide cases at their own discretion to some extent. Authorities also have plenty of scope to decide about the extent to which they punish non-cooperative behavior in less formalized settings (in the workplace, in the classroom or for any kind of self-organizing community). Our framework captures both dimensions of the typical punishment institution: the authority that applies punishment is free to decide on its extent within a set of rules that limit the severity of sanctions.

Several real-life instances exist in which different institutional procedures are used to select

executives in charge of administering justice and enforcing the law. Lay jurors are a case in point, as they are randomly selected in some countries (with a Common Law tradition), while in other countries they are appointed (most continental European countries), and yet in others they are directly elected by citizens (some cantons in Switzerland).[1] In the United States, judges at all levels of the judiciary are appointed in some states but elected in others, and this seems to influence their judicial decisions (e.g. Hanssen, 1999). The same is true for public prosecutors, with Rasmusen et al. (2009) suggesting that elections cause prosecutors to aim at higher conviction rates. At the law enforcement level, sheriffs and chiefs of police often share an overlapping mandate. Whereas the overwhelming majority of sheriffs are elected by their constituencies, all police chiefs are appointed. In general there is very little empirical evidence on how the sanctioning behavior (e.g. conviction rates, severity of penalty) of lay jurors, judges, prosecutors or law enforcement executives depends on their selection procedure. Besides the lack of data availability and severe restrictions on data use (see e.g. for data issues of jury trial data Anwar et al., 2015), a major problem for the empirical analysis is the endogeneity of the selection procedure, as different groups tend to adopt different institutional arrangements. The causal effect of endogenous institutional choice cannot be disentangled from the characteristics of the individuals who make the choice and the profile of the person in charge of administering sanctions.

Using a laboratory experiment, we investigate whether institutional endogeneity per se matters for the severity of third-party punishment in a social dilemma. After gaining experience in a multi-person prisoner's dilemma, our subjects vote whether they wish to play the same version of the game or a modification that allows for a third-party to punish defectors. The punishment decisions of an elected third-party punisher are compared to those of a third-party punisher who has been randomly appointed. We provide a theoretical framework based on the outcome-based social preference model of Charness and Rabin (2002) to explain the punishment decisions of the differently selected third-party punishers. Our experimental design allows us to control for the mentioned selection and signaling channels inherent to institutional choice settings.

For groups where the majority favors a punishment institution, we find that punishment amounts to an average of 40.2% of the maximum punishment level in the exogenous institutional setting and 14.4% in the endogenous one. The difference in punishment severity may be explained by differences in its expected effectiveness. Indeed, assigned punishment points are significantly more successful in getting defectors to contribute in the endogenous institutional setting. That is, we show that endogenous third-party sanctions are less harsh and more effective than exogenous ones, all else equal. While endogenous institutions start out generating higher

[1]For an overview of juror selection methods see (Jackson and Kovalev, 2006).

public good contributions, confirming the existence of an endogeneity premium, over time the more severe punishment implemented in the exogenous case increases contributions beyond those of the endogenous counterpart. Overall efficiency is not different across endogenous and exogenous institutions, yet the required punishment levels are significantly lower in the endogenous setting.

Our results offer an important insight for institutional choice. Punishment by a third-party is less severe when the sanctioning institution is adopted democratically, but punishment is also more effective. That is, endogenously selected sanctions are more persuasive in changing behavior than exogenously imposed ones. We further contribute to the literature by showing that voting over sanctions does not only affect the behavior of the parties who take part in the procedure, but also the decisions of the individuals who are responsible for administering them.

1.1 Related Literature

Many studies have analyzed the effectiveness of punishment institutions in enhancing and sustaining cooperation in social dilemmas (e.g. Ostrom et al., 1992; Fehr and Gächter, 2000; Andreoni et al., 2003). A burgeoning literature has explored which conditions are most conducive to cooperation, e.g. the cost-to-effectiveness ratio of punishment, group size, and whether punishment or reward systems perform better (for an overview see Chaudhuri, 2011).

Recently, several authors have investigated the effectiveness of endogenous punishment institutions, focsuing on two types of punishment regimes: *centralized formal* and *decentralized informal*. Centralized formal sanction mechanisms automatically reduce the payoff of defecting players by a certain amount. The literature has studied both costless (e.g. Tyran and Feld, 2006) and costly regimes (e.g. Putterman et al., 2011; Markussen et al., 2014). In costly regimes participants pay a fixed cost to have the scheme in place. Decentralized informal peer-to-peer punishment provides group members with the option to punish each other at a cost. Both the punishing and the punished subjects pay the cost, and typically the cost paid by the punisher is lower. Endogenously implemented centralized formal punishment regimes (Tyran and Feld, 2006; Dal Bó et al., 2010; Putterman et al., 2011; Markussen et al., 2014; Kamei et al., 2015) and decentralized informal peer-to-peer punishment regimes (Sutter et al., 2010; Markussen et al., 2014) have both proven to be more effective than their exogenous counterparts.

Tyran and Feld (2006) were the first to report on the existence of the so-called endogeneity effect, showing that cooperation in a public good game is higher when the punishment institution is enacted through a majority voting procedure as opposed to the experimenter.[2] Sutter et al.

[2]Prior to Tyran and Feld (2006), endogenous choice of institutions in collective action scenarios was studied through mechanisms other than voting. For instance, Yamagishi (1986) investigate the endogenous funding of an

(2010) confirm this regularity. In their experiment, subjects can choose whether to add a peer-to-peer sanction or reward scheme to a standard voluntary contribution mechanism (VCM). For both schemes cooperation is found to be higher when the implementation is endogenous. In Markussen et al. (2014) subjects choose between costly formal sanctions, peer-to-peer sanctions and no sanctions. With experience subjects come to prefer peer-to-peer punishment over fixed sanctions, which they manage to implement efficiently. Both sanctioning institutions are more efficient when chosen collectively by majority vote than when exogenously implemented.

The effect of selecting the punishment institution through majority voting on cooperation may result from either the endogenous process itself or from side effects that the endogenous process brings about, namely self-selection and signaling. For instance, self-selection of cooperative individuals into the same institution could account for the higher observed cooperation. Groups that implement punishment may consist of participants whose preferences differ from those that choose against punishment. In addition, the vote for the punishment institution can serve as a signaling or coordination device. That is, by voting for a certain institution participants signal their willingness to cooperate. This induces participants to infer each others' intentions from the voting outcome and to cooperate more often. While some of the previous studies discuss and partially address the signaling and self-selection issues, the seminal mechanism proposed by Dal Bó et al. (2010) manages to isolate the pure impact of endogeneity on cooperation. In their experiment, groups can vote on whether to interact in an environment with or without sanctions. The mechanism consists of a random draw that may overrule the group vote, followed by another random draw that implements one of the two environments in case the vote outcome was overruled. Controlling for self-selection through the comparison of groups that vote identically but differ on whether the choice was endogenously or exogenously implemented, the authors find a significant difference in cooperation rates. This finding is evidence of an endogeneity premium.

Kamei (2014) and Chen (2014) use the same mechanism in their experimental design and replicate this regularity. In Kamei (2014) subjects play two public good games simultaneously. A non-deterrent centralized sanction scheme can be endogenously implemented in one game, whereas a random draw exogenously implements it in the other game. He finds significant evidence of an endogeneity premium in the endogenous game and positive spillover effects to cooperation in the exogenous game. Chen (2014) investigates the endogeneity premium in an experiment where subjects vote on non-deterrent formal sanctions in the absence and presence of peer-to-peer sanctions.

---

exogenously available punishment mechanism, Ostrom et al. (1992) analyze the combined effects of communication and voting, Gürerk et al. (2006) and Nicklisch et al. (2015) allow subjects to endogenously sort into different institutions by voting with one's feet.

In a closely related line of research, several papers study features of punishment institutions that are likely to affect their perceived legitimacy, e.g. the selection procedure for punishers (Baldassarri and Grossman, 2011; Grossman and Baldassarri, 2012), the compensation of punishers (Dickson et al., 2015) and the accuracy of information available to the punisher (Dickson et al., 2009). The two former studies are closest to ours. In a lab-in-the-field experiment, Baldassarri and Grossman (2011) and Grossman and Baldassarri (2012) study a public good game with third-party punishment and compare treatments in which groups either elect the punisher or whose punisher is randomly assigned. In both treatments groups play a public good game before one group member is elected or randomly selected as punisher. In the vote treatment subjects can thus select their preferred punisher based on her previous contribution decisions. The authors find that groups with an elected punisher contribute on average more than groups with a randomly selected punisher. Their punishment decision being binary, elected and random third parties punish on average the same number of subjects per round, but differ with respect to the average maximum contribution that is punished. As punishers face different distributions of contributions, the comparison of punishment choices is however limited. In addition to differences in perceived legitimacy stemming from the punisher selection method (election vs. random draw), selection effects may play a role, i.e. in the vote treatment subjects can elect a subject as punisher with a potentially higher capability to make reasonable punishment choices.

The extant evidence suggests that the impact of endogeneity on cooperation is a behavioral regularity in several settings, while the impact on punishment is yet unclear. Our experimental study makes a novel contribution to the literature by investigating whether the implementation procedure of a third-party punishment institution per se affects the severity of the implemented punishment and the resulting cooperation patterns. We therefore employ an experimental setting with centralized punishment whose severity is chosen by a third-party and then repeatedly applied. While some existing studies allow for the endogenous choice of punishment levels, they do so in a decentralized peer-to-peer punishment setting (Markussen et al., 2014; Sutter et al., 2010). With several possible (peer) punishers it is not possible to identify the impact of endogeneity on punishment, as individuals' beliefs about the others' punishment behavior are crucial for the own punishment decision. Furthermore, as in the above-mentioned studies punishment decisions are taken in every round of the game, contribution choices and punishment choices can simultaneously affect each other. Different from Baldassarri and Grossman (2011) and Grossman and Baldassarri (2012) who focus on the selection method of a third-party punisher, we are interested in the institutional legitimacy of third-party punishment that is influenced by the implementation process and abstract from the role of personal characteristics of the punisher. Our experimental design further allows us to isolate the pure effect of endogeneity on punishment

as it controls for selection and signaling effects.

The remainder of this paper is structured as follows. Section 2 summarizes the experimental design and procedures. In Section 3 we discuss predictions for punishment and contribution behavior. Section 4 includes the results and Section 5 concludes.

2 Design

At the beginning of the experiment subjects are randomly divided into groups of four. Two different roles are assigned within a group. Three group members are A-type subjects and one is a B-type subject. The experiment consists of two parts. A-types interact in a social dilemma in both parts. After part 1, A-types decide through majority voting in what institutional setting they want to interact in part 2. Part 2 can either be identical to part 1 or modified to allow for third-party punishment, to be administered by a B-type. Subjects know that the experiment comprises two parts, but only receive instructions for the second part after the first one is completed. Types are fixed throughout the experiment, but subjects are re-matched after part 1. We employ a perfect strangers protocol such that no subject is part of the same group in parts 1 and 2. Subjects are informed beforehand that 1 of the 20 periods from each part will be randomly picked for payment at the end of the experiment. Earnings in the experiment are expressed in points, which are converted to Euro at the rate of 0.05 Euro per point. The sequence of the experiment is depicted in Figure 1.

2.1 Part 1

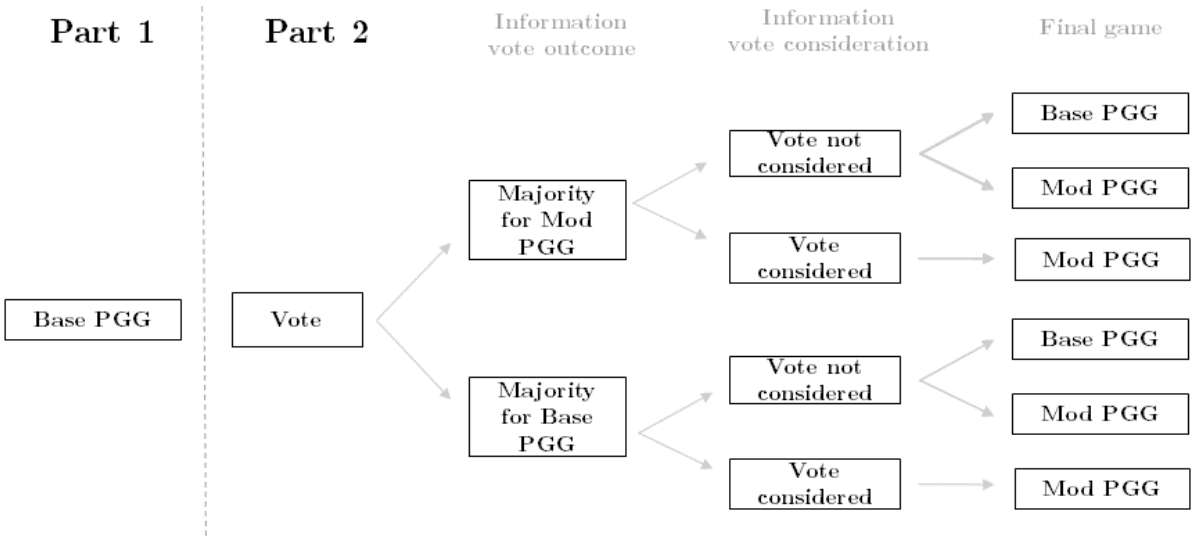
In the first part of the experiment the A-types play a 3-person prisoner's dilemma, which is equivalent to a public good game (PGG) with binary contribution choices. We stick to the latter terminology. The PGG is played for 20 periods with constant group composition. Part 1 is meant to familiarize subjects with the game and to allow them to gain experience such that they can make an informed voting decision.[3] Each A-type receives an endowment of $E_A = 70$ points in each period, which he or she can allocate to the group account ($c_i = 1$) or to the private account ($c_i = 0$). The A-types' income from the group account is the sum of all group members' contributions, $G = \sum_{i=1}^{3} c_i$, multiplied by $\alpha = 0.6$, the so-called marginal per capita return

[3]Prior research has shown that inexperienced participants prefer environments without punishment. After accumulating some experience, however, subjects' preferences reverse and punishment institutions gain support (Botelho et al., 2005; Ertan et al., 2009; Gürerk et al., 2009; Markussen et al., 2014).

(MPCR). This results in the payoff function:

$$\pi_{A_i} = E_A(1 - c_i + \alpha G) \tag{1}$$

Fig. 1: Sequence of the experiment



Notes: Only A-types participate in the Base PGG in Part 1. The dashed line between Part 1 and 2 depicts the perfect stranger rematching of groups between both parts. In Part 2 all the members of a group (the three A-types as well as the B-type) receive information on the vote outcome, vote consideration and the final game.

After each period, A-types learn the number of group members who contributed to the group account and their own period income. The B-types receive a fixed endowment of E_B=153 points (for reasons of consistency with part 2, which will be explained below). They are asked to indicate their beliefs about the A-types' behavior in the PGG. In particular, they state their expected distribution regarding the four possible outcomes in their group, i.e. in how many periods 0, 1, 2 or 3 A-types will contribute to the group account. The distribution entries must sum up to 20, the total number of periods (see Appendix A.2 for details on the belief elicitation questions). For each correct entry the B-type receives 10 points, which means that in part 1 B-types can earn up to 40 points on top of the 153 points.

2.2 Part 2

As in part 1, the three A-types interact repeatedly for 20 periods. They play either the public good game without punishment (base PGG), identical to the one in part 1, or a modified public good game with third-party punishment (modified PGG). We will first explain the modified PGG, and afterwards the voting procedure and the process that determines which of the two

games is implemented.

2.2.1 The modified PGG

In the modified PGG a third-party punishment regime is in place. Each B-type receives an endowment of $E_B = 153$ points. He or she can assign up to a maximum of 9 deduction points to each A-type who does not contribute to the group account, henceforth referred to as a 'defector'. The 153 points equal the full cooperation payoff of A-types (126 points) plus the maximum number of deduction points that can be assigned (27 points). The B-type cannot discriminate between defectors, i.e. all defectors incur the same amount of punishment. We apply the strategy method to the third-party punisher's decision and ask him or her to indicate the amount of deduction points per defector conditional on the number of defectors, which we denote by m. The vector $\mathbf{d} = (d_1, d_2, d_3)$ denotes the number of punishment points assigned by the B-type to the defectors in each of the three possible cases ($m = 1, 2, 3$). Deduction points must be an integer between 0 and 9. We follow the literature (e.g. Fehr and Fischbacher, 2004a and Fehr and Fischbacher, 2004b) in that each assigned deduction point leads to a threefold reduction of a defector's income. The resulting payoff function for the A-types is:

$$\pi_{A_i} = \begin{cases} \alpha G E_A, & \text{if } c_i = 1 \\ (1 + \alpha G)E_A - 3d_m, & \text{if } c_i = 0 \end{cases} \tag{2}$$

Assigned deduction points lead to a one-to-one reduction of the B-type's income. Punishment costs are thus small relative to the B-type's endowment. For $d_m = 0$ the function conforms to that in Equation 1 for the base PGG. The resulting payoff function for the B-type is:

$$\pi_B = E_B - m d_m. \tag{3}$$

Given that we want to study the impact of endogeneity on cooperation enforcement we exclude the possibility to punish cooperators. In our experiment the punishment institution should unambiguously serve as a tool to foster cooperation. The exclusion of 'anti-social' third-party punishment also makes the institution more attractive to cooperative A-types and is closer to real-world applications. Central authorities, e.g. judicial systems, can only punish those who violate rules. Limiting punishment to a maximum of 9 deduction points allows us to restrict it to being non-deterrent. That is, the maximum number of points that can be deducted from an A-type's period income (27=9*3) is smaller than the gain that defection brings about (28 points). That way we can make sure to have a social dilemma, which would not be the case if punishment was deterrent (i.e. higher than 9 points). In other words, with non-deterrent sanctions A-types

have no material incentive to cooperate while with deterrent sanctions cooperation would become the best response. In addition, ceiling effects could arise in such case.[4]

The punishment vector is applied throughout the entire 20 periods. In each period the actual number of defectors determines which punishment decision applies. This has the clear advantage that we can isolate the effects of the punishment decision on cooperation behavior and can exclude any endogeneity issues that would arise if contribution and punishment decisions could simultaneously affect each other. In other words, certain treatments could lead to 'back-' or 'front-loading' of punishment, e.g. an endogenously selected third-party punisher who adopts a lenient stance in the beginning but becomes harsher later on.[5]

We elicit the B-types' punishment decisions after they have received information on the vote outcome and the vote consideration, i.e. B-types know whether a majority voted in favor of the modified PGG and whether the modified PGG was introduced as a consequence of the majority vote outcome or of a random draw. While the B-types decide on the punishment vector, we elicit the corresponding punishment beliefs from the A-types (see Appendix A.2). An A-type earns 10 points for a correct belief in each of the three punishment vector entries. The modified PGG then starts. At the beginning of the second period the punishment vector is revealed to the A-types in order to avoid that uncertainty is resolved differently across treatments, as some cooperation outcomes may be more likely to occur in certain treatments. This could influence behavior in the PGG. The A-types' first period contributions are thus unaffected by the actual punishment vector or other group members' contributions. This is a deliberate design choice that allows us to assess the influence of the institution selection procedure on initial cooperation. As in the base PGG, we elicit the B-type's beliefs about A-types' behavior while they play the PGG. Furthermore the B-type is asked to fill out a questionnaire on his or her choices (see Appendix A.3). At the end of the 20 periods the B-type receives information about the A-types' contributions and the resulting payoffs in each period.

2.2.2 Voting and vote consideration

The A-types are asked to decide via majority voting whether the base PGG or the modified PGG should be implemented in part 2. After all subjects cast their vote, a random mechanism determines whether the group's vote outcome is considered. With probability $p_v = 0.5$ the votes are considered and the majority vote determines which game is implemented, leading to what we

[4] For example, Tyran and Feld (2006) observe a 93% cooperation rate under an exogenous deterrent sanction regime already.

[5] Allowing punishment to react to observed cooperation levels is certainly an interesting research question, which can be addressed by future work.

call the endogenous institutional setting (Endo). With probability $1 - p_v = 0.5$ the votes are not considered and the computer randomly decides which game is implemented, leading to what we call the exogenous institutional setting (Exo). In Exo the modified PGG is implemented with probability $p_r = 0.9$ and the base PGG is implemented with probability $1 - p_r = 0.1$. The actual probabilities are not revealed to subjects but they are aware of the procedure. All subjects, A-types as well as the B-type, learn what the majority of the A-types in their group voted for, whether the votes are considered, and which game will be implemented.[6]

The random vote overrule procedure is taken from Dal Bó et al. (2010) and makes it possible to exclude selection and signaling effects from the results. Without this procedure there would be an asymmetry between treatments, i.e. if only subjects in our Endo treatment were allowed to vote. First, a vote in favor of the modified PGG signals a preference for cooperation, which may in turn affect the B subjects' punishment behavior as well as the other A subjects' willingness to cooperate (*signaling effect*). Second, cooperative behavior after a positive vote may be attributed to a *selection effect* since groups would be composed of subjects with identical institutional preferences. In other words, those who vote for modified PGG may be more likely to contribute to the group account than those who voted for the base PGG (see Tyran and Feld, 2006 and Dal Bó et al., 2010 for a more detailed discussion). The fact that all subjects may vote, in combination with the vote overrule procedure, allows us to control for group composition effects and to keep information about A-types' preferences constant across treatments.

Within each treatment, punishment may be implemented or not. In the Endo treatments the final institutional arrangement is the one decided by the majority, so that two possible conditions may occur. In the Exo treatments, however, the opposite of what the majority voted for may be implemented. Thus, four possible Exo treatment conditions may occur. Table 1 lists all treatment conditions and the corresponding number of observations in our experiment.

We let 'P' and 'N' denote 'Punishment' and 'No punishment', respectively. Our treatment conditions are described by whether the majority vote was considered or not (Endo or Exo), whether the majority voted for the modified PGG or the base PGG (the first 'P' or 'N' after Endo or Exo), and whether the modified PGG or the base PGG was actually implemented (the second 'P' or 'N'). In ExoNP, for example, the majority voted for playing the base PGG, their vote was not considered and it was randomly determined that the modified PGG would be implemented. An intended consequence of our design is a very low number of observations in ExoNN and ExoPN, which are therefore not analyzed. This also implies that we do not analyze

[6]Groups are not informed about individual votes as this would stress the signaling content of the vote outcome, and require us to compare groups with the same vote outcome across treatments, therefore reducing the statistical power of our analysis.

Table 1: Conditions per treatment and observation numbers

Vote considered	Majority Mod PGG	Punishment	Abbreviation	A-Types	B-Types
✓	✓	✓	EndoPP	69	23
✓	×	×	EndoNN	42	14
×	✓	✓	ExoPP	69	23
×	×	×	ExoNN	3	1
×	×	✓	ExoNP	39	13
×	✓	×	ExoPN	6	2
				228	76

Notes: the number of B-types corresponds to the number of independent observations in each treatment.

the data of EndoNN, as the relevant treatment comparison would be ExoNN. Analyzing the data of EndoNN in isolation does not contribute to our understanding of sanctioning institutions as none is implemented.

2.3 Procedures

The computerized experiment was conducted at the BonnEconLab of Bonn University. Subjects were recruited on-line with hroot (Bock et al., 2014), while the software implementation was done with z-Tree (Fischbacher, 2007). A typical session lasted approximately 60 minutes and the average earnings were 13.25 Euro, including a 2 Euro show-up fee. A total of 324 subjects participated in 13 sessions (11 sessions with 24 subjects and 2 sessions with 20 subjects).[7] In order to keep instructions neutral the base PGG and the modified PGG were called "Version 1 (without deduction points)" and "Version 2 (with deduction points)", respectively.[8] In order to ensure subjects' understanding of the instructions a set of control questions was administered before the start of part 1 and another set of control questions before the start of part 2. Both parts only started when all subjects had answered them correctly. Feedback on payment (from two

[7]We ran two pilots beforehand that did not include the first part of the experiment and allowed for the punishment of cooperators. We eventually decided to change those two features in order to increase the number of groups opting for the modified PGG. In fact, most inexperienced subjects tend to prefer the simpler environment of the base PGG, which echoes much of the literature on institutional choice (see the discussion in Section 1 and footnote 3).

[8]Appendix A.1 contains a translation of the original German instructions.

randomly picked periods, one from each part) was only provided after part 2 of the experiment. At the end of the experiment subjects were asked to fill out a questionnaire that gathered their demographic characteristics (see Appendix A.3).

3 Predictions

In this section we draw on existing empirical evidence to put forward hypotheses on treatment effects for punishment and contribution choices. Further, we provide a theoretical framework that can rationalize the predicted treatment differences. Further details of the theoretical analysis, like equilibrium predictions, can be found in Appendix B.

3.1 Treatment Effects

Centralized formal punishment institutions are found to be more effective in enforcing cooperative behavior in social dilemmas like the linear public good game (Tyran and Feld, 2006, Kamei, 2014) or the prisoner's dilemma (Dal Bó et al., 2010) when they are endogenous. This means that for a given amount of punishment, cooperation is higher when the punishment institution was implemented based on the outcome of a majority vote rather than through an exogenous process. Put differently, an endogenous formal sanction is more effective in enhancing cooperation than its exogenous counterpart. In Dal Bó et al. (2010) this difference can be ascribed to the mere fact that while in the endogenous setting the institutional outcome is a result of the voting process, this is not the case in the exogenous setting. A sanctioning institution selected through majority voting may be perceived as more legitimate and can therefore trigger higher compliance vis-à-vis an exogenous institution. In particular, the direct and causal link between the voting procedure and the institutional outcome is crucial for high compliance. Whenever this link is severed, as it is the case when institutions are adopted exogenously, we can expect individuals to comply less.

Several empirical studies show that uninvolved third-parties are willing to sacrifice part of their own income to sanction non-cooperative behavior, both in one-shot and in repeated interaction (Fehr and Fischbacher, 2004a,b; Henrich et al., 2006; Kurzban et al., 2007; Almenberg et al., 2010; Engel and Zhurakhovska, 2013; Nikiforakis and Mitchell, 2014). Given that in our experiment the B-types receive identical information about the vote outcome in the ExoPP and EndoPP treatments, they should hold similar beliefs on the cooperative disposition of the A-types and choose similar punishment vectors. If punishers however anticipate the positive effect of participating in the implementation process on perceived legitimacy and cooperation, those in EndoPP may have higher beliefs about the likelihood that a given punishment level turns a defector into a cooperator as compared to ExoPP. Consequently, they would require less pun-

ishment points to reach a certain cooperation level among the A-types when the institutional process is endogenous rather than exogenous, and may therefore choose lower punishment in EndoPP.[9]

Hypothesis 1. *Punishers anticipate that punishment is more effective in enhancing cooperation if the punishment institution is endogenously implemented and therefore choose lower punishment than when the implementation is exogenous.*

In the first period of part 2 contribution decisions are yet unaffected by the implemented punishment. Controlling for the A-types' beliefs about the punishment decisions, the empirical evidence suggests that higher contributions to the public good in the first period of the game should be observed in EndoPP as compared to ExoPP. This is due to the previously discussed endogeneity premium on cooperation.

Hypothesis 2. *Controlling for the A-types' beliefs about the punishment vector, cooperation in the first period is higher if the punishment institution is endogeneously implemented (endogeneity premium).*

From the second period onwards, public good provision may depend on the extent of punishment assigned in each treatment. The harsher the implemented punishment the more subjects might contribute to the public good (see Section 3.2 for further explanation). It is ex ante unclear how the positive effect of endogeneity on cooperation will balance out with its presumably negative effect on punishment.

3.2 Theoretical Framework

We put forward a theoretical framework to illustrate how endogeneity may affect third-party punishment via legitimacy and punishment effectiveness. The analysis is restricted to a simplified stage game.[10] In the first stage A-types choose between the base PGG and the modified PGG through majority voting. In case the modified PGG is implemented, the B-type decides on a punishment vector, which specifies how many points should be deducted from defecting players for each possible number of defectors. The punishment vector is then revealed to the A-types, who subsequently make their contribution decisions. If the base PGG is implemented the B-type

[9] An alternative mechanism through which the punishment decision of the B-type may be influenced is his perceived obligation to enforce the cooperation rule by reducing the defectors' incomes. The punisher may be more willing to punish knowing that the voting of those he rules over was decisive for him being in that position, while under the exogenous institution punishers may feel less bound to spend income on punishment.

[10] First, we neglect that subjects interact repeatedly. Second, while in the experiment the steps that are described in the following are spread out over more than one period, we merge them into a one-period stage game.

do not have the option to punish and the A-types simply interact in the PGG. An equilibrium analysis of the stage game can be found in Appendix B. In this section we use the theoretical framework to explain how third-party punishment may affect A-types' contributions depending on the institutional process that introduces punishment.

We employ the outcome-based social preference model of Charness and Rabin (2002) ('CR preferences' henceforth). This model posits that individuals do not care only about their own payoff, but also about the payoff of the worst-off individual and the sum of payoffs in their group. That is, CR preferences incorporate both Rawlsian (or minimax) and efficiency (or utilitarian) concerns. The fact that the outcome-based version of CR preferences takes efficiency gains into account is important, as cooperation substantially increases social surplus in our setting.[11] CR preferences for an A type subject are expressed by:

$$\begin{aligned} &U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B) = \\ &(1-\lambda)(\pi_{A_i}) + \lambda\big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\big] \end{aligned} \tag{4}$$

where π_{A_i} and π_B are the payoffs of the A-types and the B-type as defined in Section 2, with $i = 1, 2, 3$ indexing the three A-types in the group. $\lambda \in [0, 1]$ measures how much individual i cares about the welfare of the other subjects he is matched with, and $\delta \in [0, 1]$ governs individual i's trade-off between the payoff of the worst-off individual and the maximization of social surplus. Standard preferences are nested in the model ($\lambda = 0$), but we restrict our attention to the case of $\lambda > 0$. The CR utility function for the B-type is defined accordingly.

In what follows we assume that the A-types' CR preferences are homogenous and common knowledge among A-types.[12] Unlike the standard preferences case, contributing to the public good can be an equilibrium outcome, both in the absence of a third-party punisher as in the base PGG in part 1 of the experiment, and in the modified PGG in part 2.[13] In the former case cooperation is a Nash equilibrium if enough weight is put on other individuals' welfare. The condition on λ and δ is given by the solid line in Figure 2 (cooperation is a Nash equilibrium above it). Independent of δ, the condition is met if $\lambda \geq 0.4$ for the three A-types (see Appendix

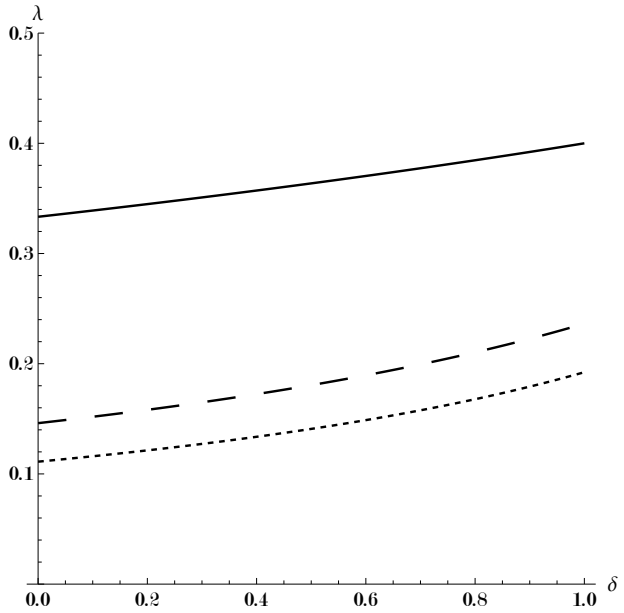
[11] The descriptive and predictive content of the Charness and Rabin model has been assessed in laboratory experiments (e.g. Daruvala, 2010; Engelmann and Strobel, 2004) and it fares well when compared to other social preference models. The model has been used to derive theoretical predictions in experimental works close to ours (e.g. Sutter et al., 2010; Markussen et al., 2014).

[12] Note that our results hinge on the specific assumptions we make about common knowledge and homogeneity of preferences. These simplifying assumptions allow us to clearly illustrate why treatment difference may occur.

[13] Under the assumptions of selfishness and rationality A-types do not contribute to the public good and B-types do not incur costs to punish defectors. As a consequence, A-types are indifferent between the base PGG and the modified PGG.

B for the derivations). To put this number in perspective, Daruvala (2010) finds an average value of $\lambda = 0.397$ in her experimental study, which means that cooperation is an equilibrium outcome for a non-negligible fraction of the population.

Fig. 2: Cooperaton and Punishment Effectiveness



Notes: Cooperation is a Nash equilibrium above the depicted lines. Each line refers to a different case: the solid line is drawn for no punishment ($d = 0$), the dashed line represents the case when $d = 5$ and $e = 1$ and the dotted line represents the case when $d = 3$ and $e = 2$.

The possibility of being punished in case of defection in the modified PGG changes the A-types' incentives to cooperate. Recall that the B-type decides on the punishment vector $\mathbf{d} = (d_1, d_2, d_3)$ using the strategy method, where the index refers to the number of defectors. Punishment is credible in our framework since the B-type decides on a binding punishment vector that is announced to the A-types before they make their contribution decisions.[14] Given that each punishment point assigned by the B-type leads to a threefold reduction of a defector's income, the B-type's decision can substantially alter the A-types' incentives. By choosing positive punishment, the B-type may successfully deter A-types from defecting. This is explained by the fact that high realized punishment sacrifices efficiency and may decrease the minimum payoff. The threat of punishment can provide the incentives for A-types with CR preferences to cooperate. The B-type implements the punishment threat as the resulting cooperation outcome

[14] In most other second- or third-party punishment games punishment is not credible because there is no incentive to punish defectors ex-post, i.e. the decision to punish is taken after the public good players have made their contribution decisions.

increases her own utility through higher efficiency and a higher minimum payoff.

To illustrate how punishment and its degree of effectiveness influence the behavior of A-types, we restrict attention to the simplest punishment vector: $d_1 = d_2 = d_3 = d$. In this case the punishment of each A-type is the same regardless of what others do. The A-types take into account the punishment vector picked by the B-type when deciding to cooperate or defect. In addition to the multiplication factor $r = 3$ that applies to each punishment point, the effect of punishment points may be magnified or attenuated by the legitimacy of the punisher who assigns them (see Section 3.1). We refer to this as the effectiveness of punishment, and denote it by the parameter e, which is a positive constant. Incorporating the effectiveness of punishment changes the A-types' utility function:

$$
\begin{aligned}
&U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) = \\
&(1-\lambda)(\pi_{A_i} + (1-e)rd) + \lambda\big[\delta \min[\pi_{A_1} + (1-e)rd, \pi_{A_2} + (1-e)rd, \pi_{A_3} + (1-e)rd, \pi_B] \\
&+ (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B + m(1-e)rd)\big]
\end{aligned} \tag{5}
$$

The effectiveness e can magnify or dampen the utility impact of punishment points. For $e = 1$ the payoff function is identical to that in Equation 2. To illustrate our point we introduce two hypothetical cases: one in which punishment is high ($d = 5$) and efficiency is low ($e = 1$) and one in which punishment is low ($d = 3$) and efficiency is high ($e = 2$). The regions above the dashed and dotted lines in Figure 2 depict the area where cooperation is a Nash equilibrium for these two cases, respectively. Comparing the zero punishment case ($d = 0$, solid line) with the high punishment case ($d = 5$ and $e = 1$, dashed line) shows that the higher the punishment, more CR preference types can be brought to cooperate. Comparing the two punishment cases we see that despite punishment being higher in the first one, the second punishment vector brings more CR preference types to cooperate due to the higher effectiveness of punishment. All else equal, if punishment is more legitimate and therefore more effective in the EndoPP treatment, more A-types will choose to cooperate as compared to ExoPP.

Important caveats apply to this illustration, namely the ad hoc nature of the effectiveness parameter and the absence of equilibrium analysis from the perspective of B-types. In Appendix B we derive equilibrium predictions for $e = 1$ assuming homogenous A-type preferences. The analysis delivers two important insights. First, if A-types have CR preferences, a third-party punisher who shares those preferences has an incentive to set high punishment. The goal is to deter A-types with mild social preferences, who would defect in the absence of punishment but cooperate when punishment is in place. This punishment strategy is deterrent in the utility-space because of CR preference subjects' efficiency and minimum payoff concerns. Second, given that the mild social preference types choose to cooperate if punishment is in place, and given

that this entails a higher payoff, we should expect them to vote in favor of the punishment institution. In other words, subjects with mild CR preferences will cooperate only if punishment is in place and they will consequently vote for the modified PGG. The intuition is that punishment acts as a commitment and coordination device for the mild CR preference types. Since highly cooperative types ($\lambda \geq 0.4$) cooperate regardless of the punishment policy, they are indifferent between punishment and no punishment. Selfish subjects ($\lambda < 0.01$) vote against punishment. In Appendix B we extend the analysis to a class of punishment vectors where deducted points can differ depending on how many A-types defect.

4 Results

We start our analysis by investigating voting behavior and its relation to public good provision in the first part of the experiment (Section 4.1). As we are mainly interested in how punishment behavior depends on the way it is put in place, in the remainder we will concentrate on those treatments in which the modified PGG is implemented (EndoPP, ExoPP and ExoNP). For the most part we will analyze behavior in ExoPP and EndoPP, the treatment conditions that offer the cleanest comparison, as in both conditions the existing punishment institution is desired by the majority of individuals. In Sub-section 4.2 we first categorize punishers in the two conditions with respect to their punishment vectors and compare punishment between treatments. We then analyze how beliefs about the A-types' cooperativeness generally influence punishment levels. In Sub-section 4.3 we compare cooperation behavior across our two main conditions and discuss efficiency implications. In Sub-section 4.4 we analyze how revealed institutional preferences interact with punishment and cooperation decisions. Here we consider behavior in ExoNP and ExoPP, as those treatment conditions only differ with respect to the outcome of the majority voting, i.e. whether the punishment institution is desired or not.

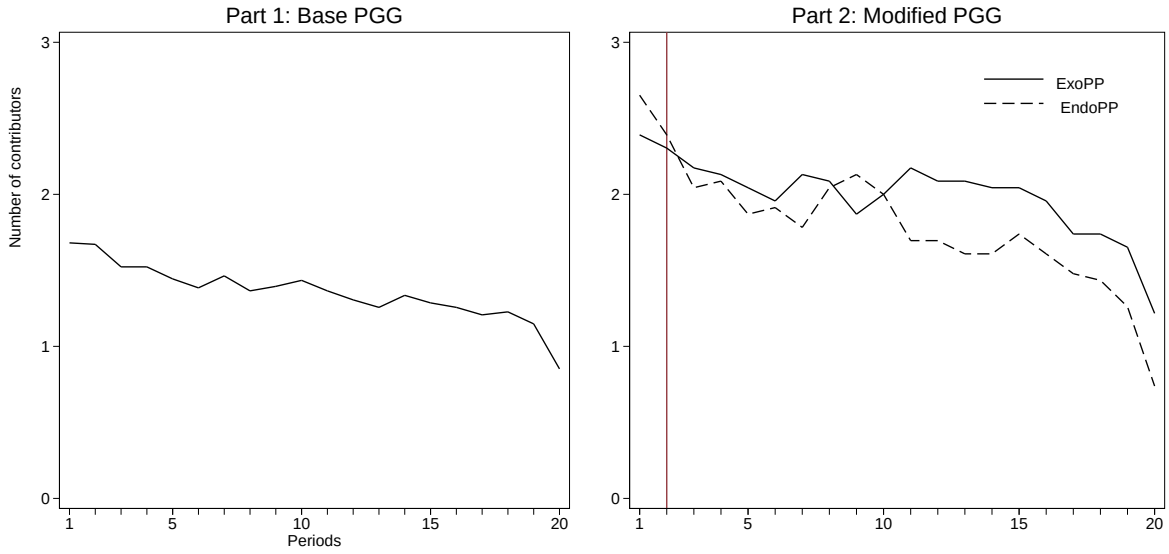
4.1 The voting decision

After interacting in the base PGG of part 1, A-types are asked to vote whether the base PGG or the modified PGG should be implemented in part 2. We find that a majority of A-types (56%) vote in favor of the modified PGG. The proportion of individuals in favor of having a non-deterrent third-party punishment institution is in line with previous comparable studies where subjects vote on the introduction of a punishment institution (see e.g. Tyran and Feld, 2006, Dal Bó et al., 2010, Markussen et al., 2014, Kamei, 2014).[15]

[15]For example, in the experiment of Tyran and Feld (2006) 50% of subjects vote in favor of having costless non-deterrent centralized punishment. Dal Bó et al. (2010) find that around 53% prefer an environment with

The left panel of Figure 3 shows the average number of contributors per group in all periods of part 1. In line with the literature on repeated PGGs we observe a steady decline in public good provision over time (see e.g. Chaudhuri, 2011). The average number of contributors starts out at 1.68 and decreases to 0.85 by the end of the 20^{th} period, with a particularly pronounced drop in the last period. In order to understand how voting behavior is influenced by public good

Fig. 3: Number of contributors



Notes: The vertical line in the right panel indicates the point in time at which A-types are informed about the punishment vector.

provision in part 1 of the experiment we estimate a logit model that relates voting behavior at the beginning of part 2 to the individual cooperative disposition, other group members' public good contributions and a measure of conditional cooperativeness. Following Gunnthorsdottir et al. (2007), among others, we use first-period contributions as a proxy for the cooperative disposition, as the contribution decision is yet unaffected by other individuals' decisions.[16] Conditional cooperativeness of subject i is measured as the average deviation of contribution behavior in t from the other two group members' (j and k) contributions in $t-1$: $\sum_{t=2}^{20} \frac{c_{i,t} - \frac{(c_{j,t-1}+c_{k,t-1})}{2}}{19} \in [-1, 1]$. The contribution of the other group members is simply the average of their contributions

costless and deterrent centralized punishment. Markussen et al. (2014) report that around 65% of subjects vote in favor of implementing costly non-deterrent centralized sanctions upon gaining experience in a social dilemma setting (around 20% of inexperienced subjects vote the same way). Kamei (2014) finds that 42% of subjects vote in favor of a costless non-deterrent informal sanctioning institution.

[16] Gunnthorsdottir et al. (2007) show in an experiment that a subject's initial contribution is an appropriate measure of their cooperative disposition.

in part 1. Table 2 reports the marginal effects of the logit estimation. We find that positive

Table 2: Determinants of voting decision

	(1)
First Period	0.24***
	(0.08)
Cond. Coop.	0.45**
	(0.19)
Contribution Others	-0.39***
	(0.14)
Observations	228

Notes: This table reports marginal effects of a logit regression model. The marginal effects are calculated at the means of covariates. *First Period* is a dummy variable for first-period contribution, *Cond. Coop.* is the conditional cooperativeness variable. *Contributions Others* includes the average of the other group members' contributions in part 1. Standard errors are clustered at the group level from Part 1 and indicated in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

experience in terms of higher average contributions of the other group members decreases the probability to vote for the modified PGG. This is intuitive as the punishment option may not be perceived as necessary to enforce cooperation. We further find that both the first period contribution and conditional cooperativeness are significantly and positively correlated with the probability of voting for the modified PGG with punishment.

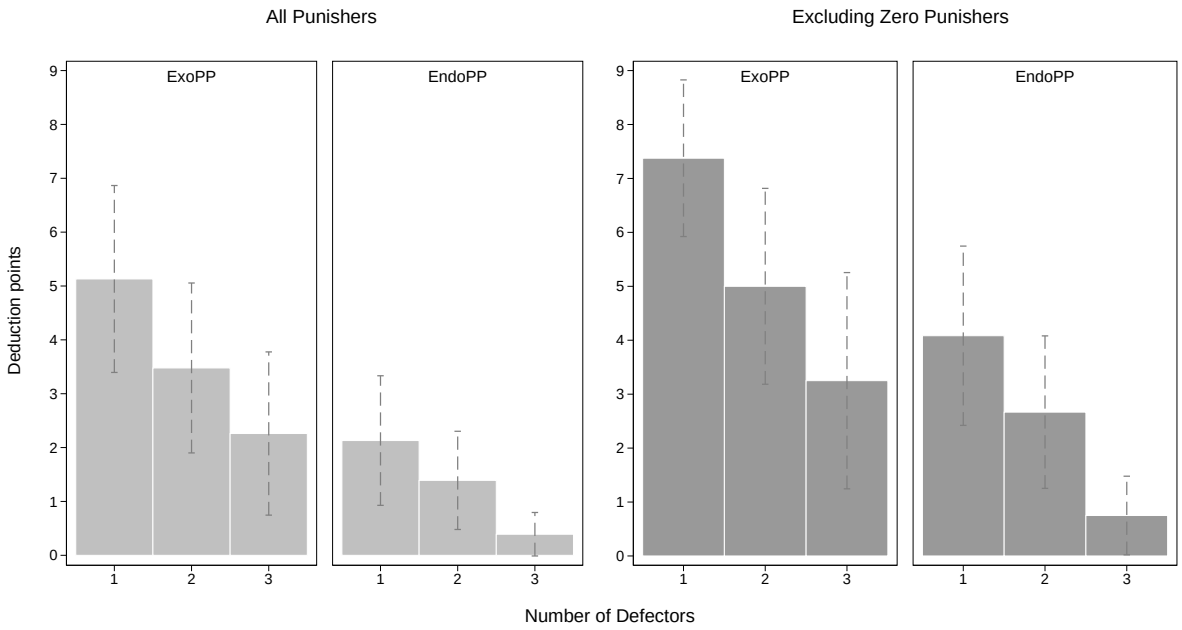
4.2 The punishment decision

4.2.1 Treatment differences and punisher types

Recall that the institutional preferences of groups in EndoPP and ExoPP are identical. What distinguishes them is whether the voting outcome was decisive or not. In EndoPP the punisher is endogenously appointed through majority vote, whereas in ExoPP she is exogenously appointed by a random mechanism. Figure 4 shows average punishment levels conditional on the number of defectors in the two treatment conditions for the complete sample of B-types (left panel) and excluding those who assign zero punishment points in all cases (right panel). Considering all B-type decisions, punishment is on average three times higher in ExoPP as compared to EndoPP. In the former condition B-types assign an average of 3.62 points while in the latter they assign 1.30 points. In line with our hypothesis, our analysis reveals that the average of all three conditional

punishment decisions is significantly higher in ExoPP than in EndoPP (two-sided Mann-Whitney test, $p = 0.03$; MW henceforth). The same is true when we consider each decision separately (MW, 3: $p = 0.07$; 2: $p = 0.07$; 1: $p = 0.01$).

Fig. 4: Punishment decisions by B-types



Notes: The dashed lines depict 95% confidence intervals.

Result 1. *Punishment is significantly higher when the punishment institution is exogenously introduced as compared to when it is endogenously adopted.*

We put forward a classification of third-party punishers with four categories based on their punishment vectors: zero' ($d_1 = d_2 = d_3 = 0$), 'conditional' ($d_1 \geq d_2 \geq d_3$), 'deontological' ($d_1 = d_2 = d_3$)' and 'others'. See Table 3 for the respective frequencies. While zero punishers are the most frequent category (39%), we find that the treatment difference is not driven by a higher number of B-types choosing zero punishment in EndoPP. Excluding those B-types from the analysis yields average deduction points of 5.21 in ExoPP and 2.5 in EndoPP, a difference that is statistically significant (MW, $p = 0.02$). When we consider each case separately, punishment is significantly different across treatments in the one defector-case, marginally significant in the two defector-case and is insignificant in the three defector-case (MW test, 3: $p = 0.13$; 2: $p = 0.09$; 1: $p = 0.00$). The second most frequent type are conditional punishers (35%). In fact, deduction points are decreasing in the number of defectors in both treatment conditions, on average, whether we exclude zero punishers or not (Jonckheere-Terpstra test, $p < 0.01$ and $p <$

0.03, respectively). A small fraction of punishers are 'deontological' (15%), while the remaining 11% constitute the 'others' category.

Table 3: Type classification of punishers

	EndoPP	ExoPP	Total
Conditional	30%	39%	35%
Deontological	4%	26%	15%
Zero	48%	30%	39%
Other	17%	4%	11%
Total	23	23	46

Notes: Numbers are rounded and therefore do not necessarily sum up to 100%.

4.2.2 The (Expected) Effectiveness of Punishment

A possible explanation for the different punishment levels observed in our two main treatment conditions is a difference in the expected effectiveness of deduction points in increasing cooperation. If third-party punishers in EndoPP expect deduction points to be more effective due to the higher perceived legitimacy of punishment, ExoPP and EndoPP punishers may implement different punishment policies.

The B-types report their beliefs about the behavior of the A-types in part 1 and part 2 of the experiment: they are asked to indicate in how many periods they think 0, 1, 2 or 3 A-types will contribute to the group account in the 20 periods. In order to test the effectiveness conjecture we relate the change in beliefs of the B-types from part 1 to part 2 to the number of assigned deduction points. As feedback on public good provision in both parts is only provided at the end of the experiment, beliefs about contribution rates in part 2 are unaffected by part 1 outcomes. However, the punishment vector is indicated before the belief elicitation, which means that part 2 beliefs reflect the B-type's punishment decision. The difference in beliefs represents the expected change in cooperation behavior as a result of the implemented punishment vector.[17]

Our variable of interest is the expected average change in cooperation per assigned punishment point. We compute it as the difference between a B-type's belief about the total number of

[17]Note that the B-types face different groups in part 1 and part 2. The changing group composition is however irrelevant for comparing changes in the punishers' beliefs between EndoPP and ExoPP, as in both treatments groups are randomly composed in part 1 and in part 2 all groups have voted with a majority in favor of the modified PGG. Other factors, like the repetition of the same game in parts 1 and 2, may also play a role in expectation formation, but these are constant across treatments.

contribution events (every time that $c_i = 1$) in part 2 and part 1, divided by the total number of deduction points that were assigned to A-types in part 2.[18] This variable provides a measurement of a punishment policy's expected effectiveness. Confirming our hypothesis, this measure is significantly higher in EndoPP as compared to ExoPP (MW, $p = 0.02$), taking an average value of 2.61 and 0.69 respectively. A third-party punisher in EndoPP believes that one deduction point leads to an increase of 2.61 cooperation events in part 2, taking part 1 as the reference point. In other words, third parties believe that a deduction point in the endogenous institution is 3.8 times more likely to increase cooperation than in the exogenous one. This analysis necessarily excludes zero punishers, as the effectiveness variable is not defined in the absence of assigned punishment points. Comparing the beliefs of zero punishers across the two treatments renders no statistical significance (MW, $p = 0.44$).

The questionnaire that the B-types answer after the belief elicitation in part 2 provides a further assessment of the effectiveness rationale. We elicit the B-types' expected effectiveness of punishment by asking them to report the probability that a defector who has been assigned one deduction point will change into contributing in the next round assuming the other two group members cooperated in the current round. In EndoPP punishers indicate a mean probability of 54%, while the corresponding percentage is 45% in ExoPP. This difference falls short of statistical significance (MW, $p = 0.30$), but underlines the higher expected effectiveness of punishment in the endogenous case. We conclude that the differences in punishment can be explained by differences in the expected effectiveness of the assigned deduction points.

A related question is whether the expected effectiveness differential materializes. We can answer this question by analyzing how the A-types respond to the number of received deduction points in part 2. Table 4 presents the estimation results of a panel model where the dependent variable is a dummy that takes the value 1 if an individual increases the contribution from the previous to the current period, and 0 otherwise. The explanatory variables are the number of received deduction points in the previous period, a treatment dummy, the interaction of the latter two variables and the other group members' average contributions in the previous period. We exclude the first period's contribution decision as subjects learned the punishment vector in the second period only. The results show that received deduction points are associated with a significant increase in next period's contribution. The interaction effect between the treatment variable and the number of deduction points is positive and significant, which means that a given amount of deduction points has a more pronounced effect on switching to cooperation when the sanctioning institution is endogenous. The fact that the B-type's punishment policy is a direct consequence of a majority decision leads her to assign fewer deduction points, which proves to be

[18] The variable has a missing value in case the B-type assigns no deduction points.

Table 4: Punishment effectiveness and cooperation

	(1)
Endo	0.58
	(0.47)
Punishment_{t-1}	0.55***
	(0.07)
Endo*Punishment_{t-1}	0.06**
	(0.03)
Contribution Others_{t-1}	-0.65***
	(0.15)
Observations	2484
Number of Groups	46
Number of Subjects	138

Notes: This table reports marginal effects calculated at the means of covariates using a logit panel model with mixed effects (including random effects at the subject level and Part 2 group level). *Endo* is a treatment dummy variable taking the value of 1 for EndoPP and 0 for ExoPP. $Punishment_{t-1}$ takes the number of received deduction points in $t-1$, $ContributionOthers_{t-1}$ takes the value of other group members' average contribution in $t-1$. The remaining variables are interaction terms. Interaction effects are calculated by the procedure proposed in Ai and Norton (2003) and Norton et al. (2004). Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

more effective vis-à-vis the deduction points assigned by a B-type who is appointed by chance.[19]

Result 2. *In line with the third-party punishers' beliefs, realized punishment is more effective in increasing cooperation among A-types when it is endogenously adopted than when the punishment institution is exogenously introduced.*

4.2.3 Cooperation beliefs

As discussed in Section 3.2 punishment may incentivize mildly cooperative subjects to contribute to the public good, whereas selfish subjects will not be affected by any punishment decision. B-types should therefore only be willing to incur punishment costs if they believe that A-types are mildly cooperative and thus susceptible to respond to punishment. We use the B-types' beliefs about the cooperation behavior of the A-types in part 1 as a proxy for a general belief about the A-types' cooperativeness and investigate how they relate to the punishment decision. We therefore estimate a regression model with the B-types' average number of deduction points as dependent variable. The independent variables are the beliefs about the number of cooperation events in part 1, a dummy for the majority vote outcome and a dummy that indicates whether the vote was overruled. Our analysis includes observations from the treatment conditions EndoPP, ExoPP and ExoNP. Estimation coefficients are presented in Table 5. We observe that cooperativeness is positively related to higher punishment. This is in line with the idea that very selfish A-types are expected to be unresponsive to punishment, and B-types therefore do not want to waste costly punishment on them. If A-types are very cooperative there is less need for punishment, whereas if they are mildly cooperative the introduction of punishment provides the right incentives for cooperation. Furthermore, the regression reveals that punishers in the overruled conditions (ExoPP and ExoNP) assign significantly higher average punishment than those in EndoPP, which confirms the non-parametric result from the comparison of ExoPP and EndoPP. The significant positive coefficient for a majority vote in favor of punishment indicates that punishers behave in line with the A-types' majority will (see Section 4.4 for a detailed discussion on the role of institutional preferences).

4.3 Public Good Provision and Efficiency

Having identified significant differences in punishment between EndoPP and ExoPP we now look at the A-types' contribution behavior in part 2 of the experiment. The average number of

[19]The results and significance levels are robust to including a random effect at the session level. Results are also qualitatively similar to a model in which the dependent variable is the first difference of contributions and robust to the inclusion of an interaction term between others' contributions and punishment points. These results are available upon request.

Table 5: Punishment decision and individual cooperation beliefs

	(1)
Expected Cooperation	0.04**
	(0.02)
Majority Vote	1.79**
	(0.88)
Overrule	1.77**
	(0.76)
Constant	-1.97*
	(1.15)
Observations	59
R-squared	0.190

Notes: Least squares regression. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

contributors in each of the 20 periods is depicted in the right panel of Figure 3.

We start by investigating first-period differences in contributions across the two treatments. Since the punishment vector is unknown at this point, only institutional differences and the A types' beliefs about the punishment vector may affect their contribution behavior. Table 6 presents the marginal effects of a logit regression with first-period contributions as dependent variable and the A-types' punishment beliefs and a treatment dummy as independent variables. We find that the probability to contribute to the public good in the first period is higher in Endo with marginal significance. This finding is in line with the existing literature on the cooperation-enhancing effect of endogenous institutions (see Dal Bó et al., 2010) as captured by our hypothesis. It is not only the information implied in the voting decision that affects cooperative behavior, but it matters whether an institution that may punish non-cooperative behavior is exogenously imposed or chosen by the affected individuals themselves.

We have shown that cooperation in EndoPP is significantly higher in the first period. After the first period, punishment decisions are revealed and implemented and may influence subsequent contribution behavior. We find no statistical significance between ExoPP and EndoPP contributions if we consider the entire 20 periods (MW, $p = 0.24$). Singling out contributions in the first 10 periods also does not produce a statistically significant difference (MW, $p = 0.85$). Cooperation levels in the two treatments seem to converge after an initial difference, possibly due to the higher punishment implemented in ExoPP. In fact, in the last 10 periods there are

Table 6: First-period contribution determinants

	(1)
Endo	0.11*
	(0.06)
Belief 3 Defectors	-0.04***
	(0.01)
Belief 2 Defectors	0.02
	(0.03)
Belief 1 Defector	0.01
	(0.02)
Observations	138

Notes: Logit model. Reported results are marginal effects calculated at the means of the covariates. Standard errors are clustered at the Part 1 group level and indicated in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1. Endo is a treatment dummy variable taking the value of 1 for EndoPP and 0 for ExoPP. BeliefDeduction'x' are A-types' punishment beliefs for the case of 'x' defectors.

on average more A-types contributing to the group account in ExoPP, i.e. when the punishment institutions has been exogenously introduced as compared to endogenously adopted. This difference is marginally significant (MW, $p = 0.09$).

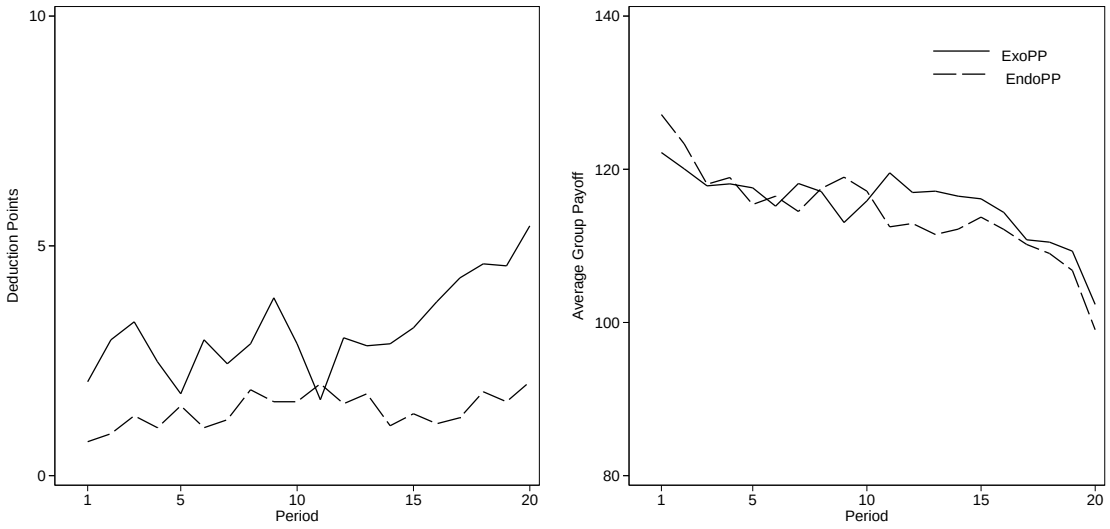
While the exogenous institution slightly outperforms the endogenous one in sustaining cooperation, this is done at the expense of higher punishment. An efficiency assessment of endogenous and exogenous institutions must take this into account. The punishment points received by the A-types in the two treatment conditions are depicted in the left panel of Figure 5. The right panel shows the average group payoff (our efficiency measure), which takes into account the punishment points deducted from the A-types and the punishment costs deducted from the B-types' endowments. We observe that punishment is higher in ExoPP, in particular towards the end of part 2, which brings the earnings in ExoPP very close to those in EndoPP. For neither player type we observe significant differences in payoffs between treatments (MW test, A: $p = 0.39$; B: $p = 0.63$).[20][21]

[20]There are no significant differences in payoffs between treatments in the first 10 or the last 10 periods (MW test, periods 1-10, A: $p = 0.90$; B: $p = 0.64$, periods 12-20, A: $p = 0.16$; B: $p = 0.75$). Our independent observation for these tests is the average payoff in a group for the A-types and the B-types respectively.

[21]The existence of the punishment institution per se does not have efficiency implications: comparing payoffs for groups where average punishment is zero we find that differences in payoffs for the A-types are insignificant

Result 3. *We find evidence for an endogeneity premium: first-period cooperation rates are higher when the punishment institution is endogenous. Overall, cooperation levels and efficiency are independent of the institution-generating process.*

Fig. 5: Punishment and efficiency across treatments



Notes: Average group payoff is defined as the average payoff of the three A-types and the B-type in each group.

4.4 The Role of Institutional Preferences

The voting outcome is public information for all subjects in a group. In order to analyze whether the punishment and cooperation decisions are influenced by the majority decision of the A-types we look at punishment decisions in the Exo treatments that differ with respect to the outcome of the majority vote: ExoNP and ExoPP. We find that average punishment is significantly lower in ExoNP as compared to ExoPP with 1.44 and 3.62 points respectively (MW test, $p = 0.07$). The regression in Table 5 echoes this finding. In the cases of one, two and three defectors, the respective average punishment levels are 1.92, 1.54 and 0.85 in ExoNP and 5.13, 3.48 and 2.26 in ExoPP. The separate case-specific analysis for different numbers of defectors yields that only the difference for the one-defector case is statistically significant (3: $p = 0.40$, 2: $p = 0.15$, 1: $p = 0.04$).[22] Our results suggest that the votes of the A-types matter for the severity of the

(MW test, $p = 1$). The B-types' payoffs do not differ across treatments as they keep their fixed endowment and do not spend money on deduction points.

[22]While the B-types' behavior is responsive to the process (Exo vs. Endo) as well as to the majority vote outcome, this is not anticipated by the A-types. Comparing the A-types' beliefs about the average punishment

punishment that is implemented. In the analysis of Section 4.1 we find that selfish A-types are more likely to vote against punishment. Thus, the vote outcome gives an indication for the cooperativeness of the A-types. In fact we find that B-types have more pessimistic beliefs about A-types' contributions in part 2 in ExoNP compared to ExoPP (average number of contribution events in ExoNP 38.77 and in ExoPP 47.48). This difference, however, falls short of statistical significance ($p = 0.13$). B-types may choose a lower punishment in ExoNP since punishment is unlikely to deter A-types and the money spent on punishment would therefore be wasted. This is reflected by the substantially larger share of zero punishers in ExoNP (54%) compared to ExoPP (30%). In addition, the punisher may simply want to respect the majority's will (no punishment) and therefore assigns few punishment points.

Comparing public good contributions between ExoNP and ExoPP reveals that contributions are significantly higher when the majority voted in favor of implementing the modified PGG with punishment (MW tests, $p \leq 0.04$), which can be explained by the selection of cooperative subjects into ExoPP.[23]

5 Conclusion

This paper investigates the role of institutional endogeneity on third-party sanctioning and the resulting consequences for cooperation behavior. A growing experimental literature on institutional choice has documented the existence of an endogeneity premium on cooperation when formal sanctioning institutions are selected through a democratic procedure. That is, groups that can choose the sanctioning institutions under which they interact tend to cooperate more. It has been shown that this phenomenon is due to the participation rights granted to groups, and not to self-selection into a preferred institution or signaling of a willingness to cooperate.

Our study compares the behavior of third-party punishers who are elected by the group she is supposed to sanction to that of third-party punishers who are appointed by chance. We show that for third-party punishment institutions endogeneity leads to milder sanctions. This result can be explained by the higher effectiveness of punishment in changing defectors' behavior in the endogenous case. A third-party punisher that is endogenously appointed anticipates that her sanctions are effective in turning defectors into cooperators, and therefore a lower level of punishment is deemed necessary. When the same institution is imposed exogenously punishment

of the B-types between ExoPP and ExoNP as well as between ExoPP and EndoPP reveals that A-types do not anticipate the B-types' consideration of the voting outcome (MW test, $p = 0.88$) or the effect of endogeneity on punishment (MW, $p = 0.64$).

[23] Our units of observations here are the average number of contributions within a group in the first 10 and last 10 periods respectively.

tends to be harsher but is not more empowered in enhancing cooperation. In spite of endogenous sanctions initially leading to more cooperation, overall the two environments exhibit identical outcomes, both in terms of cooperation and efficiency.

The idea that third-party punishers may be more lenient when an institution is endogenous to the affected individuals has previously been suggested by Feld and Frey (2002). The authors find that in cantons that score higher on a general direct democracy index (Stutzer, 1999) tax authorities impose lower maximum fines for tax evasion and lower fines in the case of self-denunciations. These implications are drawn from a context in which - unlike our setting - there is no direct link between the democratic participation rights and the task of the third-party (the tax authority), and selection and signaling effects are present. Their result resonates with our finding of lower punishment levels for defection in our endogenous institutional setting.

We can draw two main implications. First, externalities of a democratic process need to be considered when designing institutions, as individuals outside the decision process may be influenced by it. In our particular case, with an endogenous process third parties choose milder sanctions that are as effective with respect to overall efficiency as higher sanctions in the exogenous case. Second, applying an endogenous implementation process to punishment institutions may be particularly useful when punishment costs are high, as lower punishment is required to enhance cooperative behavior than when an exogenous process is applied. On the other hand, one needs to take into account that exogenous institutions may outperform their endogenous counterparts with respect to cooperation levels, as generally higher punishment levels are established. It is therefore crucial to ponder the effects that the implementation process of institutions has on the different variable features of the institutional design.

Acknowledgements

Financial support by the Max Planck Institute for Research on Collective Goods is gratefully acknowledged. We would like to thank Benjamin Bacchi, Christoph Engel, Dominik Grafenhofer, Adrian Hillenbrand, Martin Kocher, Sebastian Kube, Andreas Nicklisch, Ismael Rodriguez Lara, and Matthias Sutter for helpful comments. We are grateful to Alexander Schneeberger and Jakob Alfitian for very good research assistance.

References

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129.

Almenberg, J., Dreber, A., Apicella, C., and Rand, D. G. (2010). Third party reward and punishment: group size, efficiency and public goods. *Psychology of Punishment, Nova Publishing*.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *The American Economic Review*, 93(3):893–902.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2015). Politics in the courtroom: Political ideology and jury decision making. Technical report, National Bureau of Economic Research.

Baldassarri, D. and Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, 108(27):11023–11027.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Botelho, A., Harrison, G. W., Pinto, L., and Rutström, E. E. (2005). Social norms and social choice. Unpublished.

Charness, G., Cobo-Reyes, R., and Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization*, 68(1):18–28.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, pages 817–869.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14(1):47–83.

Chen, J. I. (2014). Obedience to rules with mild sanctions: The roles of peer punishment and voting. Unpublished.

Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *The American Economic Review*, 100(5):2205–2229.

Daruvala, D. (2010). Would the right social preference model please stand up! *Journal of Economic Behavior & Organization*, 73(2):199 – 208.

Dickson, E. S., Gordon, S. C., and Huber, G. A. (2009). Enforcement and compliance in an uncertain world: An experimental investigation. *The Journal of Politics*, 71(04):1357–1378.

Dickson, E. S., Gordon, S. C., and Huber, G. A. (2015). Institutional sources of legitimate authority: An experimental investigation. *American Journal of Political Science*, 59(1):109–127.

Engel, C. and Zhurakhovska, L. (2013). Words substitute fists: Justifying punishment in a public good experiment. Technical report, Preprints of the Max Planck Institute for Research on Collective Goods.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *The American Economic Review*, pages 857–869.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5):495–511.

Fehr, E. and Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in cognitive sciences*, 8(4):185–190.

Fehr, E. and Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

Feld, L. P. and Frey, B. S. (2002). Trust breeds trust: How taxpayers are treated. *Economics of Governance*, 3(2):87–99.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10:171–178.

Grossman, G. and Baldassarri, D. (2012). The impact of elections on cooperation: Evidence from a lab-in-the-field experiment in uganda. *American Journal of Political Science*, 56(4):964–985.

Gunnthorsdottir, A., Houser, D., and McCabe, K. (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, 62(2):304–315.

Gürerk, Ö., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312(5770):108–111.

Gürerk, Ö., Irlenbusch, B., and Rockenbach, B. (2009). Motivating teammates: The leader's choice between positive and negative incentives. *Journal of Economic Psychology*, 30(4):591–607.

Hanssen, F. A. (1999). The effect of judicial institutions on uncertainty and the rate of litigation: The election versus appointment of state judges. *The Journal of Legal Studies*, 28(1):205–232.

Harsanyi, J. C. and Selten, R. (1988). A general theory of equilibrium selection in games. *MIT Press Books*, 1.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2006). Costly punishment across human societies. *Science*, 312(5781):1767–1770.

Jackson, J. D. and Kovalev, N. P. (2006). Lay adjudication and human rights in Europe. *Colum. J. Eur. L.*, 13:83.

Kamei, K. (2014). Democracy and resilient pro-social behavioral change: An experimental study. *Available at SSRN 1756225*.

Kamei, K., Putterman, L., and Tyran, J.-R. (2015). State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics*, 18(1):38–65.

Kurzban, R., DeScioli, P., and O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human behavior*, 28(2):75–84.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., and Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences*, 111(19):6916–6921.

Markussen, T., Putterman, L., and Tyran, J.-R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *The Review of Economic Studies*, page rdt022.

Nicklisch, A., Grechenig, K., and Thöni, C. (2015). Information-sensitive leviathans–the emergence of centralized punishment. *WiSo-HH Working Paper Series*, (24).

Nikiforakis, N. and Mitchell, H. (2014). Mixing the carrots with the sticks: third party punishment and reward. *Experimental Economics*, 17(1):1–23.

Norton, E. C., Wang, H., Ai, C., et al. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, 4:154–167.

Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, 86(02):404–417.

Putterman, L., Tyran, J.-R., and Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics*, 95(9):1213–1222.

Rasmusen, E., Raghav, M., and Ramseyer, M. (2009). Convictions versus conviction rates: the prosecutor's choice. *American Law and Economics Review*, 11(1):47–78.

Stutzer, A. (1999). Demokratieindizes für die kantone der schweiz. institut für empirische wirtschaftsforschung, university of zurich. Technical report, IEW Working paper.

Sutter, M., Haigner, S., and Kocher, M. G. (2010). Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. *The Review of Economic Studies*, 77(4):1540–1566.

Tyran, J.-R. and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *The Scandinavian Journal of Economics*, 108(1):135–156.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, 51(1):110.

Appendix

A Instructions

In this appendix we present a translation of the original German instructions.

A.1 Paper Instructions

General Instructions for Participants

You are taking part in an economic experiment. Please read the following instructions carefully. You can earn money in this experiment. Your earnings depend on both your decisions and on the decisions of the other participants. At the end of the experiment, the total amount of money earned will be paid to you in cash. Additionally, you will receive a show-up fee of 2 Euro.
Throughout the experiment, monetary amounts are not quoted in Euro, but points. Your total earnings will thus be initially calculated in points. In the end the total amount of money earned during the experiment will be converted into Euro, where:

1 Point = 0.05 Euro

The experiment consists of two parts. You can earn money in both parts. So far you have received only the instructions of part 1. Instructions for part 2 will be handed out when part 1 is completed.

In this experiment there are two types of participants, A-participants and B-participants, who make different decisions. You will only get to know your own type shortly before the start of the experiment. The types will be randomly assigned. Please read the instructions about the decisions of the two types carefully.

All participants receive the same instructions. Hence, all participants receive the same information. **Talking is not permitted throughout the entire experiment.** Failure to comply will result in exclusion from the experiment and the loss of all earnings. If you have any questions, please address them to us: raise your hand and an experimenter will come to you.

On the following pages, the further course of the experiment is described in detail.

Information about the Procedure of Part 1 of the Experiment

The experiment consists of 20 periods. At the beginning of the experiment all participants will be randomly divided into **groups of 4 participants**, each group consisting of three A-participants and one B-participant. This group composition remains unchanged throughout the 20 periods. That is, you interact with the same three participants through all 20 periods.

At the end of the experiment, 1 of the 20 rounds will be randomly selected. The total amount of points earned in this period determines your payoff from part 1 of the experiment. You will not receive any information about your payoff before the end of part 2 of the experiment.

In every period, each of the three A-participants and the B-participant receive an endowment of 70 points and 153 points, respectively.

Each of the A-participants has to decide on how to allocate their endowment. There are two options:

- You choose the **private account**: Your endowment of 70 points will be allocated to the private account.
- You choose the **group account**: Your endowment of 70 points will be allocated to the group account.

The income of an A-participant is calculated differently according to the chosen account:

The **point income from the private account** directly corresponds to the amount of points allocated to it. If you allocate your endowment to the private account your income from the private account amounts to 70 points. If you allocate your endowment to the group account, your income from the private account amounts to 0 points. Nobody but yourself derives income from your private account.

Your **point income from the group account** does not solely depend on your decision, but also on the decisions of the other A-participants in your group. The point income from the group account corresponds to the sum of contributions to the group account by all three A-participants, multiplied by the factor 0.6. As soon as one of the A-participants (either you or a member of your group) chooses the group account, the group account increases by 70 points. Accordingly, the point income that every A-participant of the group receives increases by 70 x 0.6 = 42 points. The total point income of the group thereby increases by 3 x 42 = 126 points. Every A-participant of a group receives the same income from the group account, regardless of whether she contributed to the group account or not.

Depending on how the three A-participants decide to allocate their points, 4 different cases can occur: Table 1 illustrates the total group income depending on the number of A-participants who chose the group account (group-account-contributors), the number of A-participants who choose the private account (private-account-contributors) and whether an A-participant is a private- or a group-account-contributor herself, respectively.

Table 1: Decisions and Total Point Income of A-Participants

Case	Decisions of the A-Participants	Point Income of a Private-Account-Contributor	Point Income of a Group-Account-Contributor
1	3 A-participants choose the private account and 0 A-participants choose the group account	70	-
2	2 A-participants choose the private account and 1 A-participant chooses the group account	112	42
3	1 A-participant chooses the private account and 2 A-participants choose the group account	154	84
4	0 A-participants choose the private account and 3 A-participants choose the group account	-	126

As an example, we will explain case 2 of Table 1 in more detail below.

Cass 2: *Two of the three A-participants choose the private account and one chooses the group account. Hence, 1 x 70 = 70 points are in total allocated to the group account. Every A-participant receives 70 x 0,6 = 42 points from the group account. The two private-account contributors additionally receive 70 points from their private account and thus receive a total of 42 + 70 = 112 points each.*

After every period, the A-participants receive information about their point income from both the group and the private account.

While the A-participants are making their decisions, the B-participants are asked to complete a questionnaire. The corresponding instructions will be presented on the computer screen. The total point income of a B-participant equals the endowment of 153 points in every round.

The total point income of an A-participant is calculated as follows:

Point income from the private account
\+ Point income from the group account
= Total Point Income

The total point income of a B-participant is calculated as follows:

Point endowment
= Total point income

Recall: Only one of the 20 periods will be randomly selected. The total point income in this period determines your payoff from part 1 of the experiment.

Information about the Procedure of Part 2 of the Experiment

In this part, you are assigned to the same type (A-participant or B-participant) as in the first part of the experiment. The experiment consists of 20 periods. Once again, you will be randomly divided into **groups of 4 participants**. Each group consists of 3 A-participants and 1 B-participant. Your fellow group members will not be the same as in the first part of the experiment. Instead, a new group with 3 different fellow group members is formed. This grouping remains unchanged throughout the 20 periods. That is, you interact with the same three participants for all of the 20 periods.

As in part 1, 1 of the 20 rounds will be randomly selected at the end of the experiment. The total amount of points earned in this period determines your payoff from part 2 of the experiment.

In every period, each of the three A-participants and the B-participant receive an endowment of 70 points and 153 points, respectively.

As in part 1, every A-participant has to decide on how to allocate her endowment. Before heading to these decisions a vote will take place. The three A-participants vote with which of two versions of the experiment they wish the experiment to proceed (version 1 or version 2).

Version 1 - Experiment without the option to assign deduction points

In this version, the instructions remain the same as in part 1 of the experiment. The A-participants decide how to allocate their endowment. The B-participants complete a questionnaire.

Version 2 - Experiment with the option to assign deduction points

For all A-participants, the decision on how to allocate their endowment in version 2 is exactly the same as in version 1.

Additionally, the B-participant can reduce the income of the A-participants who choose the private account by **assigning deduction points**. The B-participant can also leave the income of private-account-contributors unchanged by refraining from assigning deduction points. The B-participant cannot, however, assign deduction points to group-account-contributors.

Every deduction point that a B-participant assigns to a private-account-contributor has a **deduction value** of 3 points. That is, assigning 1 deduction point reduces the private-account-contributor's income by 3 points.

Table 2: Deduction points and deduction values

Deduction points	0	1	2	3	4	5	6	7	8	9
Deduction value	0	3	6	9	12	15	18	21	24	27

Table 2 shows an overview of the resulting deduction values for all possible quantities of deduction points (0-9). If, for instance, the B-participant assigns 3 deduction points to a private-account-contributor, this leads to a deduction value of 9 points. That is, the income of the private-account-

contributor is reduced by 9 points in this round. Accordingly, if the B-participant assigns 0 deduction points to a private-account-contributor, this leads to a deduction value of 0 points. That is, the income of the private-account-contributor remains unchanged.

To each of the private-account-contributors, a maximum of 9 deduction points can be assigned. 9 deduction points lead to a deduction value of 27 points. It is not possible to assign different numbers of deduction points to particular private-account-contributors. A B-participant can assign a maximum of 27 deduction points (= 3 private contributors * 9 deduction points).

The sum of assigned deduction points to the private-account-contributors will then be deducted from the B-participant's endowment (153 points).

Hence, deduction points indicate by how many points the income of a B-participant is reduced. Deduction values indicate by how many points the income of an A-participant is reduced.

When the B-participant decides on the deduction points for the private-account-contributors, the actual decisions of the A-participants are yet unknown. Thus, decisions on the deduction points are made for the 3 possible cases when there is at least 1 private-account-contributor, i.e. independent of the yet-unknown number of private-account-contributors. The B-participant enters the deduction points for each of the three cases in table 3, which will then be presented on the computer screen. In the fourth possible case, no deduction points can be assigned, since in this case all A-participants are group-account-contributors.

Table 3: Decisions of the B-Participants

Case	Decision of the A-Participants	Deduction Points Per Private-Account-Contributor (0 – 9)
1	3 A-participants choose the private account and 0 A-participants choose the group account	
2	2 A-participants choose the private account and 1 A-participant chooses the group account	
3	1 A-participant chooses the private account and 2 A-participants choose the group account	
4	0 A-choose the private account 3A-participants choose the group account	----

As an example, we will explain case 3 of Table 3 in more detail below.

*Case 3: One A-participant chooses the private account and two A-participants choose the group account. The private-account-contributor earns 154 points and the group-account-contributors each earn 84 points (see Table 1). If, for instance, the B-participant assigns 7 deduction points to the private-account-contributor, the B-participant's endowment of 153 points is reduced by the arising cost of 7 points (153-7=146). The private-account-contributor's income is reduced by the deduction value of 3*7=21 points to 154-21=133 points. The income of both the group-account-contributors remains unchanged (84 points).*

The B-participant's decisions on the deduction points for the 3 relevant cases apply to all of the 20 periods. In each period, the deduction points determined by the B-participant apply according to the actual number of private-account-contributors. After the first period, the A-participants receive

information about the decision on the assignment of deduction points to the private-account-contributors, which the group's B-participant made for each of the 3 cases.

After the deduction point decision, the B-participant is asked to complete a questionnaire, to be presented on the computer screen. At the end of the 20 periods, the B-participant receives information about the A-participants' point allocation, the sum of deduction points assigned to the three A-participants, and their own point income in each of the periods.

The point income of an A-participant is calculated as follows:

	Point income from the private account
+	Point income from the group account
-	Deduction value (= assigned deduction points*3)
=	Total point income

The point income of a B-participant is calculated as follows:

	Point endowment
-	Sum of assigned deduction points to private-account-contributors
=	Total point income

Recall: Only one of the 20 periods will be randomly selected. The total point income in this period determines your payoff from part 2 of the experiment.

The Vote between Version and Version 2

Before the A-participants make a decision on the allocation of points, a vote takes place. The three A-participants vote on whether they wish to proceed with version 1 (without the option to assign deduction points) or with version 2 (with the option to assign deduction points) of the experiment.

After the A-participants have cast their vote, the computer randomly determines whether the vote will be considered.

- If the computer determines the vote to be **considered**, the **majority** determines whether version 1 or version 2 of the experiment applies. The assignment of deduction points to private-account-contributors is possible if the majority of the A-participants of one group (i.e. 2 or 3 A-participants) votes for this option. If only a minority (0 or 1 A-participants) votes for this option, the assignment of deduction points is not possible.

- If the computer determines the vote **not to be considered**, a **random mechanism** determines whether version 1 or version 2 of the experiment applies.

After the vote, all the group members (the three A-participants and the B-participant) will receive information about the vote result and whether it will be considered. Subsequently, all participants learn whether the option to assign deduction points to private-account-contributors will exist in the following 20 periods or not.

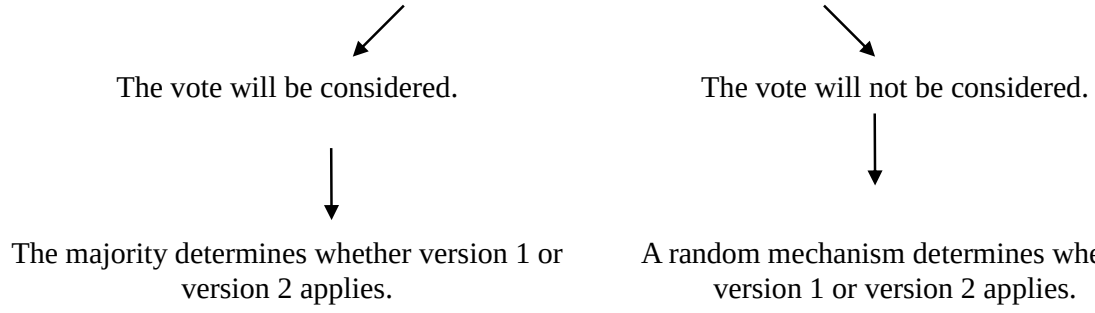
Summary

You will be divided into new groups of 4 (3 A-participants and 1 B-participants). Your fellow group members are participants with whom you have not interacted in part 1. Both group composition and your type remain unchanged for 20 periods.

A-Participants

You decide on whether you wish to proceed with version 1 or version 2 of the experiment.

The computer randomly determines whether the vote will be considered.



In both versions, you decide whether you allocate your point endowment to the private account or to the group account in each of the 20 periods. After every period, you receive information about your point income from both the private account and the group account. If deduction points can be assigned to private-account-contributors, and if you have allocated your point endowment to the private account in the respective period, you will further receive information about whether a B-participant assigned deduction points to you and, if yes, how many.

B-Participants

You receive information about the vote result in your group. You learn whether the vote result will be considered (in which case the majority determines whether version 1 or version 2 applies) or whether it will not be considered (in which case a random mechanism determines whether version 1 or version 2 applies).

- If version 2 applies, you decide on the assignment of deduction points to private-account-contributors prior to the beginning of the 20 periods. The decision on the deduction points applies according to the actual number of private-account-contributors in each of the 20 periods. Afterwards, you are asked to complete a questionnaire.
- If version 1 applies, you are asked to complete a questionnaire.

In both versions, after the A-participants have made their decisions in the 20 periods, you receive information about the A-participants' point allocation, the sum of the assigned deduction points to the private-account-contributors and about your own total point income in each of the periods.

At the end of the experiment, you will receive all payoff-relevant information from part 1 and 2 of the experiment. We kindly ask you to remain seated until you are called.

A.2 Belief Elicitation

These instructions were presented on the participant's screen.

B-type: Belief distribution on A-types' contribution behavior

The A-subjects are now deciding how to use their endowment points in each of the 20 periods. At the same time we would like to ask you to indicate your belief about how often the cases 1-4 will occur in these 20 periods. At the end of the experiment you will receive 10 points for each correct belief. If e.g. your belief about how often case 1 occurs is in line with the actual occurrence in the 20 periods, then you receive 10 points.

A-type: Belief distribution on B-types' punishment behavior

The B-subjects are now deciding on the deduction points. At the same time we would like to ask you to indicate your belief about how many deduction points the B-type assigns in the respective cases. Please indicate for each of the three fields in the table what you think the B-type entered. For each correct belief, i.e. if your belief is in line with the actual entry of the B-type in the respective field, you will receive 10 points.

A.3 Questionnaire

These questions were presented on the participant's screen.

B-type: if modified PGG is implemented

1 - Please indicate on a scale from 0 to 10 to what extent you agree with the following statements. 0=I do not agree at all, 10=I completely agree...
If I assign deduction points...

...I feel desired in my role by the A-subjects.

...I feel obligated towards the A-subjects in my role.

...I feel comfortable in my role.

2 - If I make decision that affects other people, it is important for me that they are fine with me having this decision power. [agree/disagree]

3 - Imagine the case that in a period 2 A-subjects chooses the group account and 1 A-subject chooses the private account. The B-subjects decides to assign one deduction point to the private account contributor so that his income is reduced by 3 points. How likely do you think it is that the private account contributor will choose the group account in the next period [0%, 10%, 20%, ... ,100%]?

4 - The decision of the A-subject who chose the private account is not fair [agree/disagree].

5 - I assigned deduction points in order to...

...change the behavior of the respective A-subject.

...signal that I dispraise the choice of the private account.

...reduce income differences between A-subjects.

B-type: if base PGG is implemented

1 - Please indicate on a scale from 0 to 10 to what extent you agree with the following statement. 0=I don't agree at all, 10=I completely agree...
Consider the case in which two A-subjects choose the group account and one A-subject chooses the private account in a given period. The B-subject decides to assign a deduction points to the private account contributor so that his income is reduced by 3 points.

2 - The decision of the A-subject that chose the private account is not fair [agree/disagree].

A-type: independent of which game is implemented

1 - Please explain in detail how you made your voting decision over version 1 and version 2.

2 - Under which conditions would you rather have voted for *version 2 (with deduction rule) (version 1)*? If the maximum amount of deduction points (in the experiment max 9 points)...

...would have been higher: more than 9 points.

...would have been lower: less than 9 points.

...didn't play a role for my decision.

3 - If the consequence of a deduction points for the income of the A-subject...

...would have been higher: 1 deduction point would have reduced the income by more than 3 points.

...would have been lower: 1 deduction point would have reduced the income by less than 3 points.

...didn't play a role for my decision.

A- and B-type: general questions

1 - Please indicate your gender.

2 - Please indicate your age.

3 - Please indicate your field of study.

4 - In how many experiment have you already participated?

5 - Please describe in detail to what extent you found the instructions and the experiment comprehensible. Was something difficult to understand or not clear? If yes, what was this?

B Model Predictions

In what follows we assume that both A-types and B-types have CR preferences as defined in Section 3.2. Preferences of A-types are homogenous and common knowledge to the A-types. The B-type is aware of the preference homogeneity of the A-types, but does not know the exact values of δ and λ.

B.1 Part 1

We first look at the PGG without punishment. Given that the B-type is neither influenced nor receives information on what the A-types do, we do not consider her as a player of this game. The 3 A-types, indexed as A_i (with i={1,2,3}), have to make their contribution decision in private, which consists of allocating their endowment E_A to either the group or the private account ($c_i = 1$ and $c_i = 0$, respectively). We will refer to these decisions as cooperation versus defection or contributing versus not contributing. The utility of an A-player is defined as:

$$U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}) = (1-\lambda)\pi_{A_i} + \lambda[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3})]$$

$$\text{with} \quad \pi_{A_i} = E_A(1 - c_i + \alpha G)$$

Proposition 1. *For $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)}, c_i = 1$ is a dominant strategy and full cooperation is the unique Nash equilibrium. For $\frac{1-\alpha}{2\alpha(1-\delta)} > \lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$ both full cooperation and full defection are Nash equilibria and a mixed strategy equilibrium exists. For $\lambda < \frac{1-\alpha}{2\alpha(1-\delta)+\delta}, c_i = 0$ is a dominant strategy and full defection is the unique Nash equilibrium.*

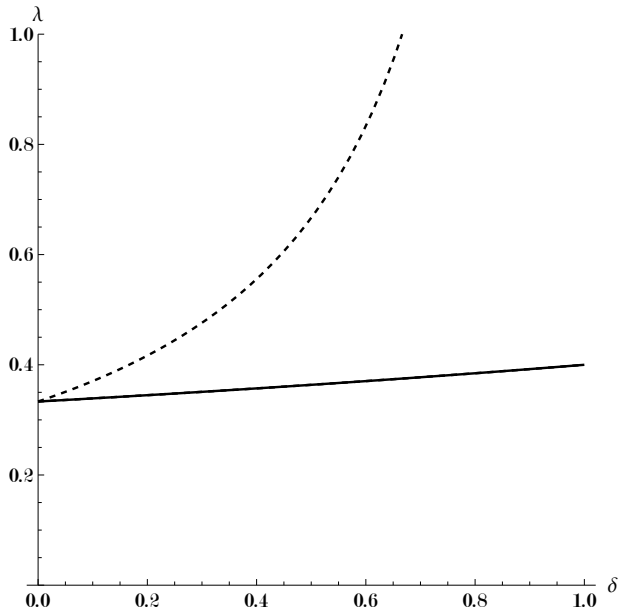
Proof. Define $U_{A_i}(c_i, c_{-i}; \lambda, \delta)$ as the utility obtained from the material payoffs associated with the contribution decision profile (c_i, c_{-i}) and parameters λ and δ. Cooperation is preferred to defection if $U_{A_i}(c_i = 1, c_{-i}; \lambda, \delta) \geq U_{A_i}(c_i = 0, c_{-i}; \lambda, \delta)$. This implies $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)}$ when two other players defect, and $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$ both when one other player cooperates and when two other players cooperate. When cooperation (defection) is a best response the corresponding Nash equilibrium follows. For values of λ such that $\frac{1-\alpha}{2\alpha(1-\delta)} > \lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$, the players cooperate if one or two others do the same, but defect when the other two defect. Full cooperation and full defection are both a Nash equilibrium, and a mixed strategy equilibrium exists. □

For the MPCR used in the experiment, $\alpha = 0.6$, the above conditions simplify to $\lambda \geq \frac{1}{3(1-\delta)}$ and $\lambda \geq \frac{2}{6-\delta}$. Figure A1 depicts these conditions.

B.2 Part 2

We now turn to the analysis of the stage game when A-types can choose between the base PGG and the modified PGG. In the first stage A-types choose between the base PGG and the modified

Fig. A1: Preference Parameters and Cooperation



Notes: The solid (dashed) line represents the second (first) condition set forth in Proposition 1. In the region above the solid line cooperation is a Nash equilibrium. It is unique above the dashed line and payoff-dominant otherwise. Below the solid line defection is the unique Nash equilibrium.

PGG through majority voting. In case the modified PGG is selected, the B-type decides on a punishment vector, which specifies how many points should be deducted from defecting players for each possible number of defectors. The punishment vector is then revealed to the A-types, who subsequently make their contribution decisions. If the base PGG was chosen no punishment option for the B-type exists and the A-types simply make their contribution decisions. We mainly focus on the case that the modified PGG is implemented and start by deriving the optimal contribution decision of the A-types given the punishment vector, and then continue with the optimal punishment decision of the B-type given the vote outcome. The game is solved by backward induction.

We define the punishment vector as $\mathbf{d} = (d_1, d_2, d_3)$, where the index indicates the number of A-types that defect. Recall that the B-type cannot discriminate between defectors in a given situation, and that it is not possible to punish cooperators. The punishment points are multiplied by a factor r before being deducted from an A-type's payoff.

B.2.1 A-type Contributions

Since the A-types' decisions have payoff consequences for the B-type, their preferences must explicitly incorporate her welfare. The utility function becomes:

$$U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) =$$
$$(1-\lambda)(\pi_{A_i}) + \lambda\big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\big]$$

$$\text{with} \quad \pi_{A_i} = \begin{cases} \alpha G E_A, & \text{if } c_i = 1 \\ (1+\alpha G)E_A - 3d_m, & \text{if } c_i = 0 \end{cases} \qquad \text{and} \quad \pi_B = E_B - m d_m$$

where m indicates the number of A-types that defect. For simplicity, we set $e = 1$ here. As discussed in Section 3, the parameter e is a measure for the effectiveness of punishment in increasing cooperation. We assume that e mirrors the perceived legitimacy of the punisher. The B-type has CR preferences identical to those of the A-types; his utility function is defined accordingly.

As in the previous sub-appendix, we start with the derivation of the A-types' best-response behavior. If two other players defect, an A-type will contribute if $\lambda \geq \frac{(1-\alpha)E_A - r d_3}{(1-\delta)[2\alpha E_A + 2(1+r)(d_3-d_2)+d_3]}$. If one other player cooperates and one other defects, or two others cooperate, an A-type player will contribute if, respectively $\lambda \geq \frac{(1-\alpha)E_A - r d_2}{\delta E_A - r d_2 + (1-\delta)[2\alpha E_A + (1+r)(2d_2-d_1)]}$ and $\lambda \geq \frac{(1-\alpha)E_A - r d_1}{E_A - r d_1 + (1-\delta)[(2\alpha-1)E_A + (1+r)d_1]}$.

For the parameter values used in the experiment ($E_A = 70$ and $r = 3$) these conditions become:

$$\lambda \geq \frac{28 - 3d_3}{(1-\delta)(84 + 9d_3 - 8d_2)} \tag{6}$$

$$\lambda \geq \frac{28 - 3d_2}{14(6-\delta) + (5-8\delta)d_2 - 4(1-\delta)d_1} \tag{7}$$

$$\lambda \geq \frac{28 - 3d_1}{14(6-\delta) + (1-4\delta)d_1} \tag{8}$$

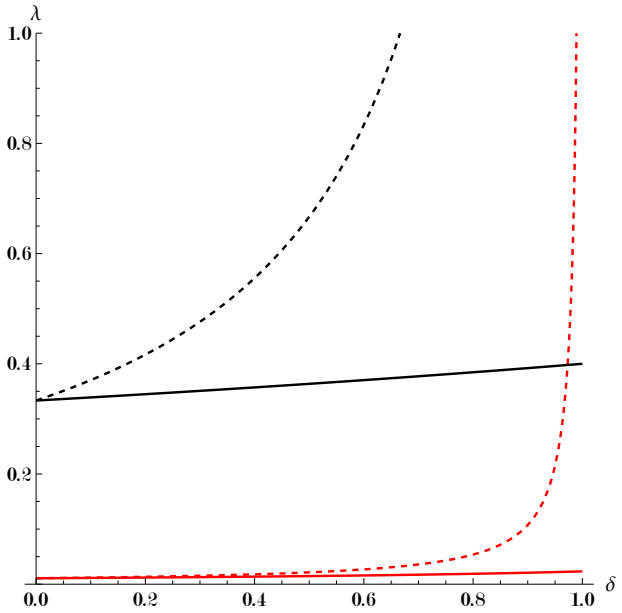
For $\mathbf{d} = (0,0,0)$ these conditions boil down to the predictions for the base PGG. The equilibrium or equilibria that result will depend on both the parameters λ and δ and the punishment vector $\mathbf{d}$. In fact, any symmetric strategy profile is an equilibrium for some combination of parameter values and punishment vector.

B.2.2 Punishment and Voting

In order to make the analysis tractable we restrict the analysis to two types of punishment vectors that represent 82% of the non-zero punishment vectors in our sample. We will start with the

simplest case: $d_1 = d_2 = d_3 = d$. This would correspond to a situation in which the B-type chooses the same level of punishment regardless of the A-types' behavior, which we refer to as a 'deontological' punishment vector in the main text. Imposing the same level of punishment for each possible outcome makes equations 7 and 8 identical, and renders equation 6 more binding than the former two as long as $d < 70/3$, which is true in our case as $d_i \in \{0, ..., 9\}$. The range of (λ, δ) for which cooperation is a Nash equilibrium is monotonically increasing in d. Figure A2 illustrates the point by plotting the equilibrium conditions for $d = 0$ and $d = 9$. The intuition is simple: increasing punishment renders cooperation a best-response for a wider range of CR preference types, as the dis-utility caused by punishment through efficiency and concerns for the lowest payoff increases.

Fig. A2: Punishment Levels and Cooperation



Notes: The black (red) lines represent the equilibrium conditions for $d = 0$ ($d = 9$).

For the remaining derivations we apply payoff dominance as an equilibrium selection criterion (Harsanyi and Selten, 1988). For example, we assume that for parameter configurations for which full cooperation and full defection are both a Nash equilibrium A-types will play the full cooperation equilibrium. The material payoff of cooperation is substantially higher than the one for defection: each A-type receives 126, compared to 70 in case of full defection. The latter payoff will be lower if punishment is positive. In the CR-utility space these differences will be more pronounced because of efficiency concerns. In addition, the fact that subjects vote in favor of punishment provides a strong signal towards coordinating on cooperation, in case both cooperation and defection are Nash equilibria. The B-type's utility function is defined as:

$$U_B(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) =$$
$$(1-\lambda)(\pi_B) + \lambda\big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\big]$$

Proposition 2. *If the B-type is assumed to choose the same level of punishment in all cases ($d_1 = d_2 = d_3 = d$), the equilibrium is characterized by the B-type setting $d^* = 9$, selfish and mostly selfish A-types ($\lambda < \frac{1}{(93-50\delta)}$) voting against punishment and all others voting in favor of punishment.*

Proof. Let $U_B(c_1, c_2, c_3, \mathbf{d}; \lambda^B, \delta^B)$ be the utility accruing to the B-type when the A-types play (c_1, c_2, c_3). She picks the punishment level $\mathbf{d} = (d, d, d)$ and has CR preferences described by λ^B and δ^B. The B-type knows that the A-types will cooperate if $\lambda \geq \frac{28-3d}{14(6-\delta)+(1-4\delta)d}$ and defect otherwise. Increasing d makes this condition less binding, i.e. full cooperation will be a Nash equilibrium for a broader range of CR preferences. We can show that $U_B(1, 1, 1, \mathbf{d}; \lambda^B, \delta^B) \geq U_B(0, 0, 0, \mathbf{d}; \lambda^B, \delta^B)$ for all $(\lambda^B, \delta^B) \in [0, 1]$:

$$153 + 378\lambda^B - 405\lambda^B\delta^B \geq 153 + 210\lambda^B - 293\lambda^B\delta^B \tag{9}$$
$$\Rightarrow \delta^B \leq 1.5 \tag{10}$$

As a result, the B-type will set $d = 9$ in order to make as many CR preference types as possible cooperate. Those A-types who would not cooperate for $d = 0$ but cooperate for $d = 9$, i.e. those for whom $\frac{2}{6-\delta} > \lambda \geq \frac{1}{93-50\delta}$, are better off in the latter case as cooperation entails a higher payoff $(70 + 293\lambda(1-\delta)$ and $126 + 405\lambda(1-\delta)$ for full defection and cooperation, respectively). Punishment does not decrease payoffs as it is merely deterrent. These A-types will thus vote in favor of punishment. A-types which are not deterred by the maximum punishment level ($\lambda < \frac{1}{93-50\delta}$) will vote against punishment. Highly cooperative A-types ($\lambda \geq \frac{2}{6-\delta}$) are indifferent between punishment and no punishment as they will cooperate in either case, and therefore are weakly in favor of punishment. □

In sum, the B-type implements a credible punishment vector that sorts A-types into their preferred institution through voting. A-types that are deterred by this punishment policy are better off under a punishment regime. Those who would be worse off under punishment do not vote for punishment.

Next, we investigate the optimal punishment vector and voting behavior under less restrictive conditions assuming a 'conditional' punishment policy. We assume that $d_1 \geq d_2 \geq d_3$, i.e. a single defector is never punished less harshly than two defectors.[24]

[24]Such a punishment policy may exist if free-riding is deemed more deserving of punishment when at least

Proposition 3. *If the B-type is assumed to choose* $\mathbf{d}$ *such that* $d_1 \geq d_2 \geq d_3$*, the equilibrium is characterized by the B-type setting* $d_1^* = 9$*, selfish and mostly selfish A-types* $(\lambda < \frac{1}{(93-50\delta)})$ *voting against punishment and all others voting in favor of punishment.*

Proof. We start by describing the equilibria that can possibly occur in the PGG. When $d_1 \geq d_2 \geq d_3$, it can be shown that both condition 6 and 7 are more binding than 8, i.e. whenever one or both of the former are binding the latter necessarily is too. However, it is not guaranteed that condition 6 is more binding than 7. One of four equilibrium configurations can be observed, depending on $\mathbf{d}$ and the A-type's (λ, δ):

- if the conditions expressed in equations 6, 7 and 8 are all binding, cooperation is the unique equilibrium.
- if equation 6 is binding (and then necessarily also equation 8) but equation 7 is not, there exist two equilibria: full cooperation and one A-type cooperating and two defecting. Since the former involves a higher payoff for all players, we select it according to our payoff dominance criterion.
- if equation 7 is binding (and then necessarily also equation 8) but equation 6 is not, there are two equilibria: full cooperation and full defection. Since the former involves a higher payoff for all players, we select it according to our payoff dominance criterion.
- in all other cases full defection is the unique equilibrium.

Given that full cooperation is an equilibrium whenever equation 8 is binding, the B-type will make it the least binding possible in order to get as many CR preference types as possible to cooperate. Since equation 8 only depends on d_1 and its partial derivative with respect to it is negative, increasing d_1 lowers the λ for which full cooperation is an equilibrium. Therefore, the B-type will set $d_1 = 9$. The remaining vector entries can take any value as long as $d_1 \geq d_2 \geq d_3$. The voting behavior of A-types is identical to what was shown in Proposition 2. □

In brief, under preference homogeneity among the A-types the B-type chooses to punish defection by one player (d_1) as harshly as possible, as this guarantees the existence of a full cooperation equilibrium for the broadest CR preference parameter range. A-types who are sufficiently cooperative, i.e. who can be persuaded to cooperate by this punishment level, vote in favor of

one other A-type cooperates. Equivalently, such a policy can be rooted in the fact that it is easier to bring one defecting A-type to cooperate than achieving the same when all A-types defect. This punishment strategy would also be picked by a B-type who wanted to keep expenditure relatively constant across the three possible cases (recall that d_1 has to be paid once while d_3 has to be paid three times). In fact, the average punishment vector in the experiment conforms to $d_1 \geq d_2 \geq d_3$.

punishment. All others vote against punishment. The punishment of two and three defectors (d_2 and d_3) does not play a role in this result as long as we impose the payoff-dominance selection criterion. Relaxing this assumption (e.g. by allowing for mixed strategy equilibria) would allow us to say more about the second and third punishment vector entries, but a meaningful analysis would also require a detailed distribution of the B-types' beliefs on λ and δ.