Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2013/9

Deterrence by Imperfect Sanctions – A Public Good Experiment

Christoph Engel

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2013/9

Deterrence by Imperfect Sanctions
A Public Good Experiment

Christoph Engel

May 2013

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

Deterrence by Imperfect Sanctions
A Public Good Experiment*

Christoph Engel

Abstract

Sanctions are often so weak that a money maximizing individual would not be deterred. In this paper I show that they may nonetheless serve a forward looking purpose if sufficiently many individuals are averse against advantageous inequity. Using the Fehr/Schmidt model (QJE 1999) I define three alternative channels: (a) identical preferences are common knowledge, but inequity is not pronounced enough to sustain cooperation; (b) heterogeneous preferences are common knowledge; (c) there is preference uncertainty. In a linear public good with punishment meted out by a disinterested participant, I test two implications of the model: (a) participants increase contributions in reaction to imperfect punishment; (b) imperfect punishment helps sustain cooperation if participants experience free-riding.

Keywords: Deterrence, imperfect sanction, inequity aversion, Fehr/Schmidt preferences, public good experiment, centralized punishement

JEL: C91, D03, D63, H41, K13, K14, K42

* Help by Lilia Zhurakhovska in designing and running the experiment and helpful comments by Michael Kurschilgen and Aniol Llorente-Saguer on an earlier version are gratefully acknowledged.

1. Introduction

Sanctions are frequently imperfect. Crime often goes unnoticed. The police do often not have enough resources to investigate petty crime. Criminal sanctions are rarely so severe that the expected value of committing crime becomes negative. Likewise, as a rule tort only entitles the victim to compensation. If there is only a small risk that the victim will not sue, or will not win in court,[1] the expected loss from being sued is below the expected gain from tortious behavior. In this paper I model and experimentally test one reason why imperfect sanctions might not be pointless: a sufficient fraction of the addressees might hold social preferences. Sanctions might help stabilize the willingness of inequity-averse individuals to do what is in society's best interest.

In the US, deterrence of criminal behavior is not only a desirable side-effect of the application and enforcement of criminal law; it is its official purpose (18 U.S.C. § 3553 (a) (2) (B)). Private law liability is also interpreted as a deterrent (for a review see Schwartz 1994). Nonetheless, criminal law and liability law frequently do not deter a money maximizing individual. In most parts of the US, prostitution is a crime. Yet the risk for a prostitute to be arrested has been estimated to be as low as 1:450 (Levitt and Venkatesh 2007**:** 5). In 2007, in the US 58.75% of all persons under correctional supervision were on probation, while only 20.70% of them actually did time (Glaze and Bonczar 2007**:** 2); arguably the disutility from fulfilling probation requirements is considerably smaller than disutility from incarceration, and likely also below utility from the original crime. In torts, double or treble damages are the exception.[2]

Psychological effects make the threat with sanctions even less powerful. It has been shown that many criminals grossly overestimate the profitability of their trades, some of them expecting the criminal act to be 10 times as profitable as it actually is (Wilson and Abrahamse 1992**:** 368). Also criminals tend to be risk seeking (Arneklev et al. 1993**;** Cochran et al. 1994**;** Esbensen and Deschenes 1998**;** Winfree and Bernat 1998**;** LaGrange and Silverman 1999**;** De Li 2004**;** Kerley et al. 2008), which implies that the disutility from the sanction threat is below its expected value.

Of course, imperfect sanctions are not pointless. Criminal sanctions are also meant as acts of retribution, they may incapacitate the perpetrator, or they may help the defendant become a worthy member of society (18 U.S.C. § 3553 (a) (2)). Private law liability at least compensates those plaintiffs who sue and win for the damage caused by defendant. Yet arguably, incomplete sanctions at best partly deter undesired behavior, if at all. Therefore seemingly the forward looking purpose of sanctions, their governance effect, is severely weakened.

1 And if the perpetrator's gain is a mirror image of the victim's loss.

2 The most important cases are antitrust (15 U.S.C. § 15) patent (35 U.S.C. § 284) and false claims against the government (31 U.S.C. § 3729(a)(1)(G)).

This argument however presupposes that all threatened with sanctions are equal, and that they all maximize gains from crime, or from other socially undesirable behavior. Intuitively, this seems implausible. For a host of reasons, individuals may not simply compare the gain from crime with the expected loss from a sanction. They may expect additional social sanctions, like scorn from their peers. They may have moral compunctions. They may consider a criminal act as detrimental to self-esteem. They may be risk averse and therefore weigh the prospect of the sanction more heavily than its expected value. Or they may hold preferences such that the loss for the victim of the crime reduces their utility.

In this paper I focus on the last explanation. I do not mean to say that alternative explanations are immaterial. All I want to show is that social preferences provide one consistent explanation for the governance effect of imperfect sanction. Sanctions that would be too weak to deter individuals who straightforwardly maximize their utility from non-social preferences may suffice to deter individuals holding social preferences, even if their disutility from outperforming others is small. Imperfect sanctions extend the domain of cooperation to individuals who do not care strongly about others, but who are also not immune to the detriment they inflict on others. For such individuals, even imperfect sanctions deter.

I proceed in two steps. I develop a simple model, derived from the canonical formalization of social preferences in Fehr and Schmidt (1999). To test the implications of the model for the governance effect of imperfect sanctions, I use one experimental standard design that has been at the origin of this theoretical model, the linear public good. The only difference from the standard design is the implementation of punishment. While this literature has normally used decentralized punishment (see only Fehr and Gächter 2000**;** Balliet et al. 2011**;** Chaudhuri 2011), I entrust an additional anonymous participant with the right to punish the four active members of the current group.

I prefer centralized over decentral punishment to exclude a confound with positive utility another active group member may derive from sanctioning a free-rider, like revenge (cf. Masclet et al. 2003**;** Crosetto et al. 2012). I prefer sanctions meted out by an experimental participant over automatic sanctions executed by the computer for two reasons. First the source of uncertainty then is the behavior of another human agent, not a deterministic machine. This is closer to the source of uncertainty in the field. Second, I may not only study reactions of would-be perpetrators to imperfect sanctions. I may also study the decision of authorities to content themselves with imperfect sanctions where the design of the experiment would have made it easy for them to make the sanctions perfect.

With considerable frequency my experimental authorities indeed leave sanctions imperfect, meaning that the loss in profit resulting from the sanction is smaller than the benefit from exploiting other group members. This gives me variation in my first explanatory variable. I also administer a standard test of social preferences (Liebrand and McClintock 1988) and find them to be fairly heterogeneous. I find that participants who have disutility from advantageous inequity increase contributions in reaction to punishment even if punishment had not

been deterrent. This even holds if participants have made the experience that others completely free-ride on their efforts to provide the public good. Imperfect sanctions do thus indeed extend the domain of cooperation if participants hold social preferences.

Testing these theoretical expectations in the lab has the usual advantage of clean and controlled conditions. In my case, a second advantage is equally important. In the field it would already be difficult to measure the severity of the sanction. While formal sanctions are usually well documented, the informal, social correlates are much harder to observe. In the lab, I exclude them through anonymous interaction. Due to the notorious dark-field problem, reliable estimates of the risk of criminal sanctions are even more difficult to generate. In the lab, I can not only observe how often which type of socially undesirable behavior attracts which sanction. I even have complete control over the experiences each individual active member has made with the sanctioning authorities. Most importantly, rare cases like the prostitution example mentioned above notwithstanding, in the field gains from crime or deviant behavior are at best guesswork. In contrast, in the lab I can precisely quantify gains. For these reasons I believe that the inevitable loss in external validity resulting from going to the lab is justified.

The remainder of the paper is organized as follows. Section 2 presents the design of the experiment. Section 3 relates it to the experimental literature. Section 4 develops the theoretical model and derives hypotheses. Section 5 reports results. Section 6 concludes with discussion.

2. Design

I conduct a linear public good experiment with the standard payoff function

$\pi_i = e - c_i + \mu \sum_{n=1}^{N} c_n$	(1)

where e is the endowment, c_i is the contribution of this player, $0 < \mu < 1 < N\mu$ is marginal per capita rate, n is generic for any player, player i included, and N is group size. In the experiment $e = 20\ Taler, \mu = \frac{4}{10}, N = 4$. At the end of the experiment, each Taler is converted into .04 €.

I randomly assign a fifth player to each group. This player earns a fixed amount of 1 € (the equivalent of 25 Taler). She receives 20 tokens that she may use for punishing any of the active players. Each punishment point assigned destroys three Taler of the active player's income. Any punishment point the authority does not use is credited with .01 €. If the authority does not use any punishment points, she thus earns 1.20 €, the equivalent of 30 Taler. I implement a fine to fee ratio of 1:12.

Interaction is anonymous. After the end of the first round, active group members are informed about the contributions of all other active members, and of punishment points assigned to

each of them, if any. They also learn their period income as well as their current total income from the experiment.

After the end of the first round, there is a surprise restart with another 10 rounds of the same game. The purpose of the separate one shot experiment is to provide me with a manipulation check. If social preferences explain cooperation in this first phase of the experiment, I can be sure that results are not driven by the prospect of gains in future rounds. In the second phase of the experiment, participants learn that they will be rematched every round, but that roles are kept constant. I choose a stranger design for reasons of internal and external validity. In terms of internal validity, the stranger design excludes that cooperation is driven by reputation effects. Moreover in the legal application that has triggered this research, judges are unlikely to regularly meet the same defendants. Following the procedure that is standard in the experimental literature (see e.g. Charness 2000**;** Montero et al. 2008), I assign participants to matching groups of 10, but do only tell them they will be re-matched every period, not that matching groups have limited size. This procedure is meant to guarantee independent observations, without inducing participants to try to second guess group composition. Each period, one authority and four active players are randomly matched.

While I need variance in punishment policies, it would not be good for my explanatory variable if punishment was chaotic. In the interest of inducing an exogenous norm, at the end of the instructions I therefore inform participants about mean contributions in a structurally similar experiment of myself with a co-author in the same lab (Engel and Irlenbusch 2010). I communicate this information in the form of a graph.[3]

My theory expects participants to hold heterogeneous preferences. I use a standard tool from social psychology to measure inequity aversion, the social value ring measure (Liebrand and McClintock 1988).[4] This test has participants define a series of allocations between themselves and another anonymous participant. Aggregating these choices I learn their willingness to pay for either not exploiting or, to the contrary, for exploiting others.[5]

The experiment was conducted in the Cologne Laboratory for Economic Research in 2012. The experiment is programmed in zTree (Fischbacher 2007). Participants were invited using the software ORSEE (Greiner 2004). 90 student participants of various majors had mean age 25.39. 44.44 % were female. Participants on average earned 15.11 € (19.82 $ on the days of the experiment), 14.80€ for active players, and 16.38 € for authorities. I had 3 sessions of 30 participants.

3 For details see the transcription of the instructions in the appendix. This way, I also follow the lead of Tyran and Feld (2006), who also exogenously induced a contribution norm.

4 See the results sections for the precise mapping of social value orientation to inequity aversion.

5 Since my results might also be explained by the individual specific degree of risk aversion, I also administer the standard test by (Holt and Laury 2002). Since this variable turns out uninformative, I do not report tests using it.

3. Related Literature

Tyran and Feld (2006) also investigate deterrence by imperfect sanctions in a public good. They exogenously impose a norm of full contribution. In their exogenous treatments, this norm is either not sanctioned at all, it is enforced by a deterrent sanction, or by a mild sanction that would not deter a money maximizing agent. In their endogenous mild treatment, group members can vote for mild sanctions. In their endogenous severe treatment, they can vote for severe sanctions. They do not find a significant effect of exogenously imposed mild sanctions, while mild sanctions chosen by majority vote have a beneficial effect. They explain the difference by a commitment effect, which translates into a higher willingness of conditional cooperators to make substantial contributions.

My paper mainly differs in the following respects. I do find a beneficial effect of exogenously imposed imperfect sanctions. I have a formal behavioral model, and offer inequity aversion as an alternative explanation. Based on this model, I also test in which ways the beneficial effect of non-deterrent sanctions hinges on the experienced heterogeneity of the population. In the conclusions I discuss further features of my design that might explain why I find results that differ from the earlier paper. Yet explaining this difference is not the purpose of this paper. Rather I want to test the theoretical expectation that imperfect sanctions should discipline inequity-averse participants.

There is, of course, a rich literature on the effects of a punishment option on contributions in a linear public good (for recent summaries see Balliet, Mulder et al. 2011**;** Chaudhuri 2011). This literature shows that contributions are sensitive to manipulations of the severity of punishment (Casari 2005**;** Egas and Riedl 2008**;** Nikiforakis and Normann 2008**;** Ambrus and Greiner 2011) and of the certainty of punishment (Grechenig et al. 2010**;** Sousa 2010). Punishers react to the opportunity cost of punishment (Carpenter 2007).

Most publications on public good experiments do not derive their hypotheses about the effect of punishment from formal behavioral theory. In principle, a case could be made for punishment reacting to perceived intentions (for models of intentions see Rabin 1993**;** Dufwenberg and Kirchsteiger 2004**;** Falk and Fischbacher 2006) (from an evolutionary perspective see Carpenter et al. 2004), to violations of exogenous norms (Andreoni and Bernheim 2009**;** Dufwenberg et al. 2011), or to violations of efficiency (Charness and Rabin 2002**;** Engelmann and Strobel 2004). Yet to the extent the effects of punishment have been modeled, all papers have assumed that punishees are motivated by inequity aversion (Fehr and Schmidt 1999**;** Bolton and Ockenfels 2000).

Thöni (2011) uses inequity aversion to explain antisocial punishment. Gürerk et al. (2010) hypothesize that punishment may not explain a cooperative equilibrium if players are inequity averse and the setting is such that the current ability to contribute to the public good is a fraction of total earnings from the game. Kosfeld et al. (2009) use inequity aversion to explain

under which conditions players will endogenously introduce a punishment option (also see Masclet and Villeval 2008).

Closest to my approach is the theoretical paper that has started the literature (Fehr and Schmidt 1999). They show that, if sufficiently many group members are sufficiently intensely inequity averse and if there is a punishment option, there may exist equilibria where the strongly inequity averse players contribute positive amounts while the remaining players contribute nothing. Yet they are interested in explaining how the second order public good (Yamagishi 1986**;** Heckathorn 1989) is overcome: if inequity averse individuals suffer enough disutility from being exploited, they gain by punishing freeriders. By contrast I am interested in understanding how an imperfect threat with sanctions serves a purpose.

From a different angle, Johnson et al. (2009) is related. They also combine a public good with punishment option with a second game that measures preferences for income equality. Yet they do not use the social value ring measure for the purpose and, more importantly, they are not interested in explaining sensitivity to punishment, but the willingness of active group members to engage in peer punishment. They find willingness to punish to be the more pronounced the more an individual cares about equality.

4. Model and Hypotheses

Each period, (first stage) payoff is given by (1). Since $N\mu > 1$, all players contributing their entire endowment would be efficient. Yet since $\mu < 1$, keeping the endowment is the best response for a player who holds standard preferences. Consequently, if all players hold standard preferences, all of them keeping their entire endowment prescribes the unique Nash equilibrium of the game. Authorities may keep any punishment point they do not use. Therefore money maximizing authorities do not punish at all. If preferences are common knowledge, this is anticipated by active players. Hence we have a first prediction:

P_1: If all participants hold standard preferences and this is common knowledge, active players do not contribute to the public good. Authorities do not punish.

Given earlier public good experiments (for overviews see Ledyard 1995**;** Zelmer 2003**;** Chaudhuri 2011), and given a companion paper on the behavior of authorities (Engel and Zhurakhovska 2012), I am sure to reject **P_1**. One reason for a different outcome would be that all active players hold Fehr-Schmidt preferences (Fehr and Schmidt 1999), with defined and identical parameters α and β, and all of this being common knowledge. Then payoff would result from (1), but utility would be given by (2):

$u_{it} = \pi_{it} - \frac{1}{3}\alpha \sum_{j \neq i} \max\{\pi_{jt} - \pi_{it}, 0\} - \frac{1}{3}\beta \sum_{j \neq i} \max\{\pi_{it} - \pi_{jt}, 0\}$	(2)

We still have the uncooperative equilibrium since, if all contribute nothing, both the second and third terms in (2) are 0, so that we are back to (1). For a cooperative equilibrium in pure strategies to exist, the critical condition is

$$\begin{aligned} & u_{it}(c_{it}, c_{it}, c_{it}, c_{it}) > u_{it}(0, c_{it}, c_{it}, c_{it}) \\ & \Leftrightarrow e - c_{it} + 4\mu c_{it} > e + 3\mu c_{it} - \beta(e + 3\mu c_{it} - [e - c_{it} + 3\mu c_{it}]) \\ & \Leftrightarrow \beta > 1 - \mu \end{aligned} \qquad (3)$$

or with the parameters of the experiment $\beta > \frac{3}{5}$. This is high, but not unheard of: Fehr and Schmidt (1999**:** 844) assume $\beta \leq .6$, yet Blanco et al. (2011**:** Fig.1) also find considerably higher values. Since the utility function is linear in the decision variable c_{it}, we must have a corner solution. While any contribution level is an equilibrium, it seems plausible that a group of inequity averse players coordinate on the efficient outcome; of course risk dominance would still predict that players coordinate on the uncooperative equilibrium.

Next allow for preference heterogeneity, but keep the assumption that preferences are common knowledge. Assume that one of four group members holds standard preferences, while the remaining three hold Fehr/Schmidt preferences, with identical parameters. In this setting, players holding social preferences would contribute their entire endowments if the following condition holds:

$$\begin{aligned} & u_{it}(c_{it}, 0, c_{it}, c_{it}) > u_{it}(0,0, c_{it}, c_{it}) \\ & \Leftrightarrow e - c_{it} + 3\mu c_{it} - \frac{1}{3}\alpha(e + 3\mu c_{it} - [e - c_{it} + 3\mu c_{it}]) \\ & > e + 2\mu c_{it} - \frac{1}{3}\beta * \frac{2}{1}(e + 2\mu c_{it} - [e - c_{it} + 2\mu c_{it}]) \\ & \Leftrightarrow \beta > \frac{3}{2}(1 - \mu) + \frac{1}{2}\alpha \end{aligned} \qquad (4)$$

or, with the parameters from the experiment, $\beta > \frac{9}{10} + \frac{1}{2}\alpha$. This is implausibly high. *A fortiori*, no cooperation is predicted if, in a group of four, a participant who holds social preferences à la Fehr/Schmidt herself only expects one more group member to hold the same preferences, while she expects the remaining two group members to maximize profit.

Let us next relax the common knowledge assumption. Now individuals must rely on their beliefs. Of course if all believe all others to hold Fehr/Schmidt preferences with defined and identical parameters, and if all believe all others to believe that they hold such preferences, the conditions for a cooperative equilibrium established in (3) hold; despite the objective uncertainty players decide as if it was common knowledge that all hold identical Fehr/Schmidt preferences. Yet cooperation already becomes more difficult to sustain if only one player ("Player 2") believes one other player ("Player 1") to only hold Fehr/Schmidt preferences with probability $p < 1$. Now Player 2 plays the game of figure 1**Fehler! Verweisquelle konnte nicht gefunden werden.**. Nature assigns a type to Player 1. Conditional on her type, this player chooses be-

tween keeping her endowment ("0") and contributing ("e").[6] Yet Player 2 only observes Player 1's choice, not her type. She therefore does not know in which branch of the game tree she is.

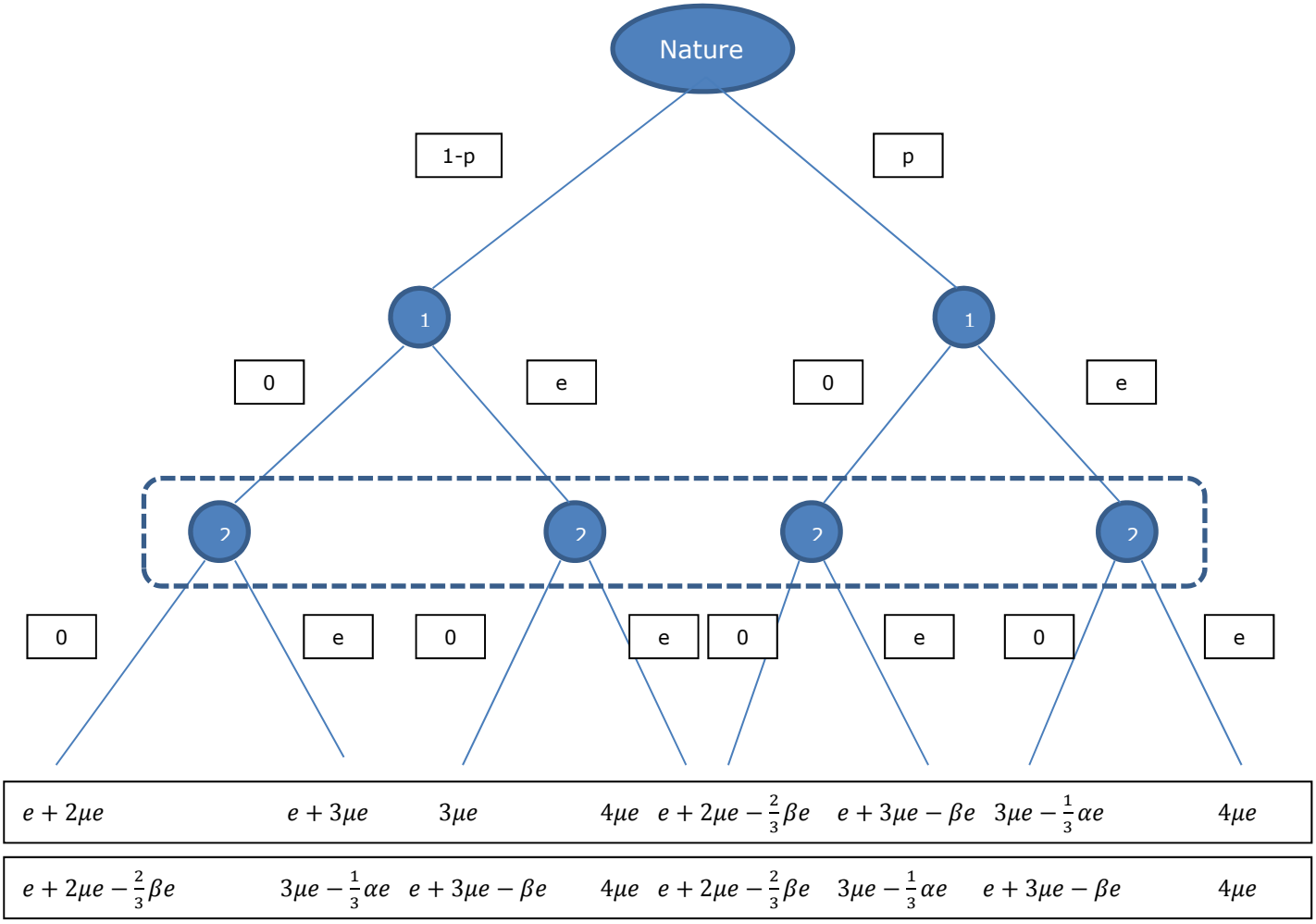


Figure 1
Bayesian Model

The critical comparisons are $e + 3\mu e > 4\mu e > e + 3\mu e - \beta e$ and $e + 2\mu e - \frac{2}{3}\beta e > 3\mu e - \frac{1}{3}\alpha e$. The first set of parameters makes sure that defection is the best response to all other participants contributing their endowments if the player is selfish, while contributing the endowment is the best response if the player holds Fehr-Schmidt preferences. The second comparison implies that even a player holding Fehr-Schmidt preferences prefers not to contribute if the other player does not contribute as well. With these assumptions, the player who is uncertain may calculate her expected utility in case she defects, and in case she contributes her endowment. If she expects the first player to defect as well, obviously defection is her best response; the first equilibrium in pure strategies is given by $s_1(0,0), s_2(0)$, where s_i stands for the strategy of player i. In case the first player plays strategy $s_1(0, e)$, the second player compares expected utility, given either of her choices. She cooperates if

6 It is straightforward to extend Figure 1 and the ensuing calculations to any $c_{it} \leq e$. I refrain from doing so since the figure then becomes difficult to read.

$$u_{it}(e) > u_{it}(0) \Leftrightarrow (1-p)\left(3\mu e - \frac{1}{3}\alpha e\right) + p4\mu e > (1-p)\left(e + 2\mu e - \frac{2}{3}\beta e\right) + p(e + 3\mu e - \beta e) \quad (5)$$

The cooperative equilibrium in pure strategies requires that

$$p > \frac{3(1-\mu) + \alpha - 2\beta}{\alpha + \beta} \quad (6)$$

Following the same logic, I could extend this to uncertainty about the type of more than one active player. This approach assumes a common prior p. From the second period on, active players would have to update this prior, using Bayes rule.

These considerations give us a second prediction:

P_2: a) If all players are sufficiently averse against advantageous inequity and this is common knowledge, all of them contributing their entire endowments is an equilibrium.
b) If one sufficiently inequity averse player is uncertain whether one other player is sufficiently averse against advantageous inequity, the inequity averse player may still cooperate herself provided she deems the probability of defection sufficiently small, or provided she is sufficiently averse against advantageous inequity.
c) If three of four players are averse against advantageous inequity, but believe a fourth player to be selfish, they defect for all plausible degrees of inequity aversion.

Let us now introduce sanctions meted out by an authority. We first assume that all active players hold standard preferences, and that the authority has credibly committed to inflicting a sanction s proportional to $e - c_{it}$ to any active player who contributes less than her entire endowment. This changes each active player's decision problem to

$$\pi_{it} = \begin{cases} e - c_{it} + \mu \sum_{n=1}^{N} c_{nt} & if \quad c_{it} = e \\ e - c_{it} + \mu \sum_{n=1}^{N} c_{nt} - s(e - c_{it}) & if \quad c_{it} < e \end{cases} \quad (7)$$

Contributing her entire endowment becomes the best response if $s > (1-\mu)$ or, with the parameters of the experiment, if the sanction is more severe than 12 tokens for a player who has contributed nothing. Note that the authority can destroy 60 tokens in total. The authority thus has power to completely deter free-riding, even if all group members were to contribute zero tokens to the public good.

This brings us to the research question of this paper: how do social preferences and sanctions interact? Let us first reconsider (3): all active members hold identical Fehr/Schmidt prefer-

ences and this is common knowledge. The fact that the authority has committed to sanction deviations from full contributions changes the active players' problem to the following:

$$
\begin{aligned}
& u_{it}(c_{it}, c_{it}, c_{it}, c_{it}) > u_{it}(0, c_{it}, c_{it}, c_{it}) \\
& \Leftrightarrow e - c_{it} + 4\mu c_{it} > e + 3\mu c_{it} - \beta(e + 3\mu c_{it} - [e - c_{it} + 3\mu c_{it}]) - se \\
& \Leftrightarrow \beta > \frac{(1-\mu)c_{it} - se}{c_{it}}
\end{aligned}
\qquad (8)
$$

Comparing with (3) we see that aversion against advantageous inequity must be less pronounced. If inequity averse players consider coordinating on full contributions, the condition simplifies to $\beta > 1 - \mu - s$. The more the sanction is severe, the less the demands on aversion against advantageous inequity. Sanctions and inequity aversion are substitutes. The more the sanction is severe, the smaller the degree of aversion against advantageous inequity that is needed for sustaining cooperation. More importantly for the purposes of this paper, the reverse also holds true: the more participants are averse against advantageous inequity, the less sanctions must be severe to sustain cooperation. Sanctions that would be imperfect for selfish participants still serve a purpose if all participants hold social preferences. Sanctions extend the domain of inequity aversion. A degree of inequity aversion that would be insufficient in and of itself to maintain cooperation may suffice if it is combined with mild, seemingly imperfect sanctions.

From a policy perspective, the next step is even more important. Let us reconsider the situation where one of four active group members holds standard preferences while the remaining are averse against inequity. If there is a credible threat with sanctions, the problem of inequity averse participants expressed in (4) changes to

$$
\begin{aligned}
& u_{it}(c_{it}, 0, c_{it}, c_{it}) > u_{it}(0,0, c_{it}, c_{it}) \\
& \Leftrightarrow e - c_{it} + 3\mu c_{it} - \frac{1}{3}\alpha(e + 3\mu c_{it} - [e - c_{it} + 3\mu c_{it}]) \\
& > e + 2\mu c_{it} - \frac{1}{3}\beta * 2(e + 2\mu c_{it} - [e - c_{it} + 2\mu c_{it}]) - se \\
& \Leftrightarrow \beta > \frac{3}{2}(1-\mu) + \frac{1}{2}\alpha - \frac{3}{2}\frac{se}{c_{it}}
\end{aligned}
\qquad (9)
$$

Compared with (4) the condition has a third term. Now it becomes easier for the inequity-averse participants to sustain cooperation among themselves, even if they expect one participant to free-ride. Most importantly, the sanction serves a purpose even if it is not severe enough to deter the one participant holding standard preferences. It helps inequity-averse participants overcome the disutility of being exploited by the one selfish participant.

Finally introduce sanctions into the game of Figure 1. Obviously, with $s > 12$, even a selfish player would cooperate. The interesting case is $12 > s > 0$. Such an imperfect sanction does not deter a selfish player. If this player expects the other player to hold Fehr/Schmidt preferences, sanctions are not necessary to make the former player contribute. The same holds true

if her β is large enough, or her subjective p is small enough, to induce her to cooperate anyway. Imperfect sanctions become critical, though, if the uncertainty is so pronounced that this player's degree of aversion against advantageous inequity is too small to induce her to cooperate, given the perceived risk of one other player being selfish. In that constellation, it becomes relevant that sanctions change (5) to

$$u_{it}(e) > u_{it}(0) \Leftrightarrow (1-p)\left(3\mu e - \frac{1}{3}\alpha e\right) + p4\mu e > (1-p)\left(e + 2\mu e - \frac{2}{3}\beta e\right) + p(e + 3\mu e - \beta e - se) \quad (10)$$

Proceeding the same way as before, this changes (6) to

$$p > \frac{3(1-\mu) + \alpha - 2\beta}{\alpha + \beta + 3s} \quad (11)$$

There is an additional term $3s$ in the denominator that makes demands on optimism less stringent.

Inequity aversion thus predicts three socially beneficial effects of imperfect sanctions:

P3: a) Sanctions that are too weak to deter a selfish player may help a group of inequity-averse participants sustain cooperation even if the degree of inequity aversion in and of itself would not be sufficient to bring cooperation about.
b) Imperfect sanctions extend the willingness of inequity-averse participants to tolerate exploitation by free-riders.
c) Imperfect sanctions require a smaller degree of optimism if one player who is inequity averse herself is uncertain whether another player is selfish.7

I cannot directly test these hypotheses. Whenever there are social preferences, there are multiple equilibria. From studying choices I can therefore not distinguish whether one of the predictions **P_2** or **P_3** is refuted, or whether participants expected other participants to coordinate on another equilibrium. Preferences are not common knowledge. This makes it impossible to directly test **P_2** and **P_3.** I could have changed the design. I could have run the ring measure test first, and could in every round have informed participants about the ring measure scores of the remaining three members of their current group. Yet that would have changed the research question. That way, I would have seen the effects of making differences in social value orientation salient. Participants would have been induced to decide whether to be consistent between allocation decisions and decisions involving a risk of exploitation. Similar arguments speak against eliciting beliefs about other group members' choices or social value orientation in each and every round. I would have seen the effects of forcing participants to construct a mental map of the group they are interacting with.

7 In the Appendix, I show in which ways authorities react if they anticipate active players' aversion against advantageous inequity.

I instead have a design that allows me to test what I am interested in: the effects of imperfect sanctions. Specifically by my design I hold strategic uncertainty constant. Preferences are not common knowledge and participants are rematched every period so that the degree of strategic uncertainty resulting from the heterogeneity of preferences does not diminish over time. Through entrusting punishment decisions to an additional participant, I aim at inducing heterogeneity of punishment and specifically expect some of the sanctions to be imperfect. In this framework I test the following two hypotheses:

H1: In a linear public good, imperfect sanctions induce participants to increase contributions if participants are averse to advantageous inequity.

H2: This even holds if participants have been paired with complete freeriders.

5. Results

My main question is whether punishment is instrumental even if it is imperfect. For this test I need panel data. Yet I start with manipulation checks.

a) Manipulation Checks

The left panel of Figure 2 collects social value orientation scores of active participants. The distribution follows the typical pattern. The mode is at 0. Such participants behave like the agents of economic textbooks. They are exclusively interested in payoff for themselves. A small number of participants are rivalistic. The have a negative score. This implies that they are willing to give up some money for themselves to reduce the payoff of their anonymous partners even more, and thus increase their relative payoff. The majority however have some willingness to pay for improving the payoff of their anonymous partners, and thereby reduce the difference in payoffs. However only very few participants have a score of 45 or more. A score of 45 results if a participant equalizes payoffs to the maximum extent possible. An even higher score results if a participant is even willing to increase her partner's payoff if this means a smaller payoff for herself.

In the Fehr/Schmidt model an agent is indifferent between an increase in her own payoff and a reduction of advantageous inequity if she has $\beta = 1$. If she is selfish, she has $\beta = 0$.[8] In the model the lower bound for cooperation to be a best response is $\beta = \frac{3}{5}$. This translates into a social value score of 27. In the experiment, 87.5% of all active players had a score below this most optimistic benchmark. We thus have enough variance of social value orientation to use

8 The ring measure is more encompassing than the Fehr/Schmidt model. The Fehr/Schmidt model allows for $\beta > 1$. But since there is a max-operator, this does not imply that the active player has a preference for making the passive player even better off than herself. It just means that she strongly prefers equal payoffs over having a higher payoff herself. By contrast, the ring measure allows for scores above 45. They imply that a player has a willingness to pay for making an anonymous partner better off even if this means a smaller payoff for herself than for the other player.

the social value score as an explanatory variable, and enough scores below the lowest possible benchmark to make imperfect sanctions meaningful.

The right panel of Figure 2 demonstrates the effect of social value orientation on the contributions in the public good when there is no shadow of the future. As one sees, the effect is pronounced, and in the expected direction. The visual impression is supported by statistical analysis (Tobit, explaining contributions in the first round with the social value score, cons 10.219 (.953), social value score .174 (.047), N = 72, p model .0003). I can thus exclude that the effect requires repeated interaction, and thus a shadow of the future.

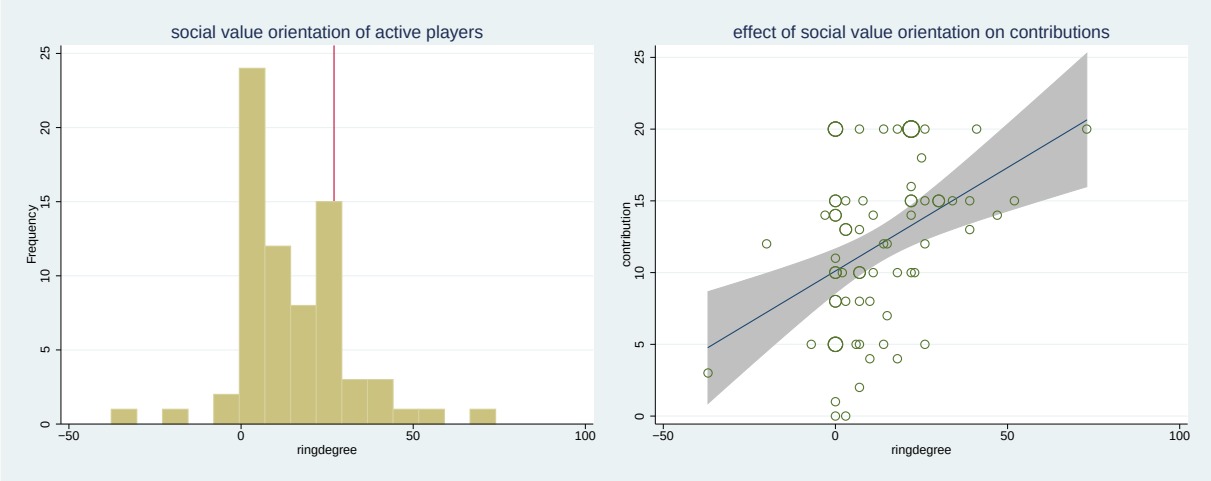


Figure 2
First Round Choices of Active Players
left panel: red line is at ringdegree = 27, equivalent of $\beta = \frac{3}{5}$
right panel: bubble size reflects frequency, line is best linear approximation, with 95% confidence interval

Finally Figure 3 shows that punishment authorities that have punished at all have sometimes meted out fairly harsh sanctions. Yet in almost half of the cases where there was punishment, it was imperfect (102 vs. 152 cases).[9] If I include cases where authorities have not punished an active participant at all, deterrent sanctions are only inflicted in less than 20% of all cases (152 of 792). We thus also have enough variance of punishment and, most importantly, enough non-deterrent sanctions.

9 Punishment deters a selfish participant if $s > 12 - .6 * contr$.

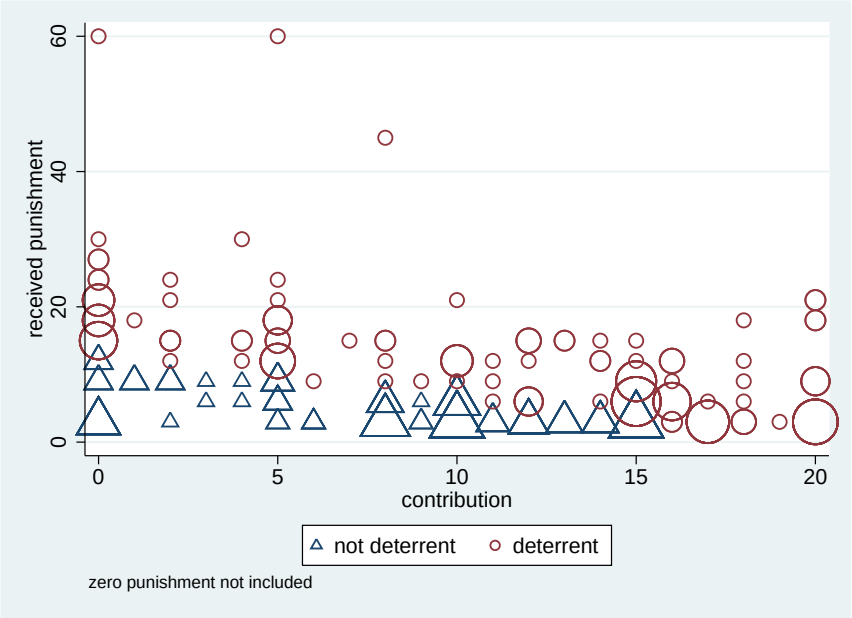


Figure 3
Punishment
bubble size indicates frequency

b) Effect of Imperfect Sanctions

I now turn to my research question, the effect of imperfect sanctions. Model 1 of Table 1 replicates a standard finding from public goods with punishment:[10] the effect of punishment on the level of contributions in the subsequent period is negative. This seemingly surprising result has a simple explanation. Those who receive punishment do not immediately adjust their choices to the mean contributions of the remaining group members. Yet as Models 2 to 4 demonstrate, punishment is not pointless. If I analyze first differences, i.e. changes in contributions over time, I see the expected positive effect. Those who have been punished contribute significantly more in the next period (Model 2). Model 3 shows: even if I control for the fact that punishment was deterrent, and interact the level of punishment with the dummy for punishment having been deterrent, the main effect of the level of punishment remains significant. Actually it even becomes stronger. In this specification, the main effect captures the effectiveness of non-deterrent punishment. I therefore conclude

> *Result 1: Punishment induces players in a linear public good to increase contributions in the subsequent period even if punishment is not severe enough to deter a participant who maximizes her payoff.*

10 In all models, standard errors are clustered at the level of matching groups, taking the possibility into account that observations might be contaminated by earlier experiences, despite the fact that groups are rematched every period.

	Model 1	Model 2	Model 3	Model 4
dv	levels	first differences		
lagged punishment	-0.091 0.063	0.230*** 0.038	0.366** 0.106	0.442* 0.135
punishment was deterrent	3.331** 0.685		1.094 1.084	2.559^{+} 1.217
lpun*ldeterr	-0.086 0.064		-0.191 0.124	-0.377* 0.144
lpun*svo	-0.020^{+} 0.009			-0.015 0.009
ldeterr*svo	-0.057* 0.02			-0.070* 0.026
lpun*ldeterr*svo	0.025* 0.009			0.020^{+} 0.009
cons	13.024*** 0.15	-0.598*** 0.11	-0.761** 0.173	-0.751** 0.159
N	720	720	720	720
individuals	72	72	72	72
clusters	9	9	9	9
R^2 between	0.061	0.001	0.002	0
R^2 within	0.072	0.14	0.149	0.171
R^2 overall	0.044	0.101	0.104	0.123

Table 1
Deterrent Effect of Imperfect Sanctions

standard errors in parenthesis

linear with individual fixed effects, se clustered for 9 matching groups

lpun: lagged amount of punishment, ldeterr: punishment was deterrent, svo: ring measure score

*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

Model 4 tests $\mathbf{H_1}$. In that hypothesis I had not only expected imperfect sanctions to be instrumental, but I had expected this effect to be driven by participants' aversion to advantageous inequity. Since the Hausman test is significant, I must estimate fixed effects models. Since an individual's social value score does not change over time, I do not estimate the main effect of this score. But my statistical model identifies all interactions with this variable. This suffices for my research question.

Controlling for social value orientation is important for seeing the effects of punishment. In Model 4 I now find a weakly significant main effect of punishment being deterrent, in the expected direction: participants adjust their contributions more intensely if punishment had been deterrent. Yet note the significant negative interaction. The main effect of punishment being

deterrent is neutralized through the interaction effect if punishment had been more severe than 6.79 tokens.[11] Very severe sanctions do not work better.

However the most important finding for my research question is the significant negative interaction between the deterrence dummy and social value orientation. The more an individual is averse to advantageous inequity, the less it matters whether a sanction is severe enough to even deter a selfish participant. This is precisely what I had hypothesized in $\mathbf{H_1}$.

Figure 4 casts further light on this finding. The figure collects marginal effects of a one unit increase in punishment on the change of contributions in the subsequent period, split up by the degree of aversion against advantageous inequity. If punishment is deterrent, it has a significant effect on behavior, if only social value orientation is positive. The more social value orientation is pronounced, the more intensely a participant reacts to an increase in the severity of punishment. If punishment is deterrent, severity and social value orientation are complements.

Interestingly, the opposite holds if punishment is not deterrent. The model even predicts a decrease in contributions if punishment increases and the target had pronounced aversion against advantageous inequity, yet with high social value orientation the marginal effect is not significantly different from zero. I do, however, find a significant marginal effect of imperfect sanctions if social value orientation is positive, but not strong. Imperfect sanctions thus precisely work for that type of participants where my theory expected them to matter: participants whose aversion against advantageous inequity is too weak to sustain cooperation without punishment.

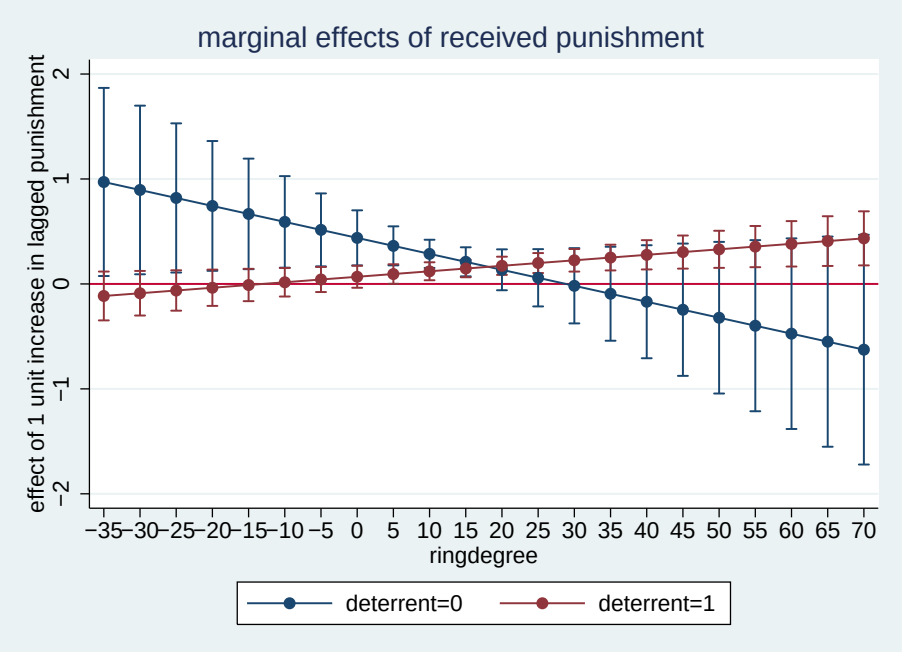


Figure 4
Marginal Effects from Model 4 of Table 1

11 2.559/.377 = 6.788.

This yields

> *Result 2: Imperfect sanctions work best for individuals with positive, but weak aversion against advantageous inequity.*

In my second hypothesis $\mathbf{H_2}$, I expect that non-deterrent sanctions may even help sustain cooperation if inequity-averse participants, in the previous period, had the experience of interacting with complete freeriders. To test this hypothesis, I need a richer statistical model. In Table 2 I repeat the final model from Table 1, but now also control for the number of freeriders met in the previous period[12], plus all interaction effects.

All coefficients of regressors that did already feature in Model 4 of Table 1 look similar. The model that does not control for the number of freeriders in the previous period therefore captures the situation well where there actually was no freerider, which is the reference category of the model in Table 2. The critical coefficient for $\mathbf{H_2}$ is the net effect of lagged punishment + lagged punishment*#freeriders. Since the model controls for punishment being deterrent and the interaction of this dummy with the severity of punishment, this net effect captures whether imperfect punishment deters even if there was one complete freerider. The critical net effect is smaller than the main effect of lagged punishment, but still significantly different from zero (coef = .337, p = .0107). Actually, the net effect is about the same even if there is more than one freerider and remains significant (coef = .385, p = .0172).

This gives me

> *Result 3: Imperfect sanctions even help sustain cooperation in the presence of complete freeriders.*

lagged punishment	0.573** 0.165
punishment was deterrent	3.346^{+} 1.462
lpun*ldeterr	-0.623** 0.165
lpun*svo	-0.050*** 0.007
ldeterr*svo	-0.058 0.093
lpun*ldeterr*svo	0.055*** 0.006
one freerider	0.563 0.469
more than one freerider	-0.719 0.612

12 In line with my theory, I only consider the freeriding of *other* group members.

lpun*1fr	-0.236^{+} 0.109
lpun*2fr	-0.188 0.168
ldeterr*1fr	-1.996 1.64
ldeterr*2fr	-1.728 2.112
lpun*ldeterr*1fr	0.418* 0.148
lpun*ldeterr*2fr	0.367 0.255
svo*1fr	-0.119^{+} 0.062
svo*2fr	-0.109 0.068
lpun*svo*1fr	0.043** 0.009
lpun*svo*2fr	0.043** 0.009
ldeterr*svo*1fr	-0.019 0.106
ldeterr*svo*2fr	-0.027 0.093
lpun*ldeterr*svo*1fr	-0.037** 0.008
lpun*ldeterr*svo*2fr	-0.046** 0.011
cons	0.708 0.664
N	720
individuals	72
clusters	9
R^2 between	0.013
R^2 within	0.237
R^2 overall	0.092

Table 2

Deterrent Effect of Imperfect Sanctions Conditional on Number of Freeriders

standard errors in parenthesis

linear with individual fixed effects, se clustered for 9 matching groups

lpun: lagged amount of punishment, ldeterr: punishment was deterrent, svo: ring measure score,

1fr: one freerider in previous period, 2fr: more than one freerider in previous period

*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

In the final step, I investigate in which ways inequity aversion is critical for this result. Figure 5 collects marginal effects for an increase in severity by one token, conditional on the number of free riders, whether punishment was deterrent, and the degree of inequity aversion. The left panel shows that the number of freeriders is indeed critical. If the participant, in the previous period, has not experienced complete freeriding by another participant, deterrent punishment has little effect. A one unit increase in severity then never significantly affects in which ways the participant adjusts her contributions. By contrast the reaction to non-deterrent sanctions is pronounced. They have a beneficial effect only if this participant is not very sensitive to disadvantageous inequity. The maximum social value orientation score at which the effect of higher severity is still (significantly) positive is 5.

By contrast if, in the previous period, the participant has met with one freerider (middle panel), the result looks similar to Figure 4, i.e. similar to a model that does not control for the number of freeriders. Whenever the participant has a positive social value score, deterrent sanctions have a beneficial effect, the more so the more the participant cares about other participants' payoff. Inequity aversion and deterrent sanctions are complements. By contrast, inequity aversion and non-deterrent sanctions are substitutes. An increase in severity only has a beneficial effect if social value orientation is not too pronounced (score of 10 or less).

Interestingly, if the participant had to face more than one complete freerider, deterrent sanctions no longer work, however strongly the participant cares about other group members' well-being. By contrast, imperfect sanctions are effective for participants who are mildly averse against advantageous inequity (social value score of 15 or less).

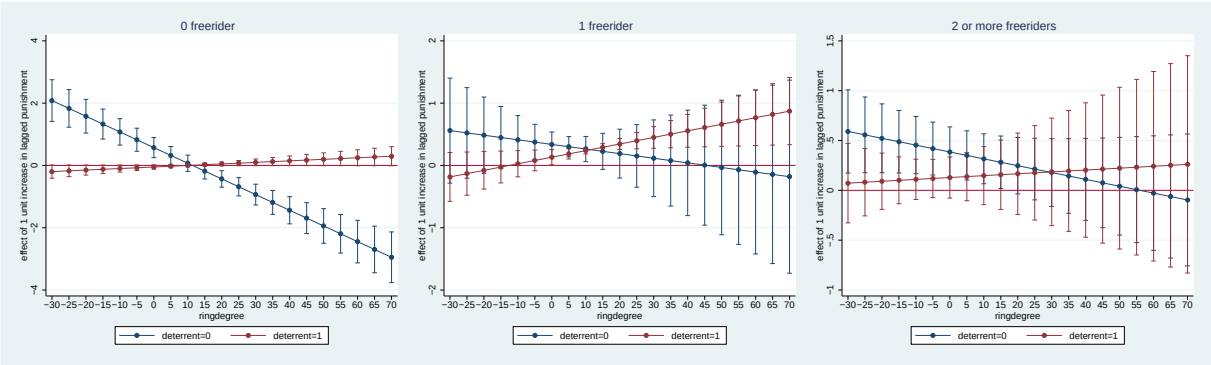


Figure 5
Marginal Effects from Model in Table 2
effect of a 1 unit increase in lagged punishment
blue line: punishment was not deterrent, red line: punishment was deterrent

This leads to

> *Result 4: a) In the absence of any complete freeriders or if there is more than one complete freerider, deterrent sanctions are not instrumental. Non-deterrent sanctions only have a beneficial effect if the participant is not too strongly averse against advantageous inequity.*
> *b) In the presence of one complete freerider, deterrent sanctions and inequity aversion are complements, non-deterrent sanctions and inequity aversion are substitutes.*

6. Conclusion

As a matter of fact, sanctions are often imperfect. The expected value of the sanction, i.e. the probability of enforcement multiplied by the severity of the sanction, is often smaller than the expected benefit from crime or tort. Sometimes the legal order cannot avail itself of a sufficiently severe sanction. Take the person who ignores the rules for access to donor organs and does so to save her own life. Even if the legal order does not bar capital punishment in the first place, buying a donor organ will likely not qualify. In other instances, society cannot afford perfect sanctions. Many crimes go undetected because the police have not enough personnel to investigate them. Prisons are costly, which is one of the reasons for granting probation. Finally, the legal order may dislike deterrent sanctions for other normative reasons. It for instance is afraid that sending first offenders to jail will introduce them to a criminal career.

From a rational choice perspective, deterrent sanctions are appealing. If only there is a credible threat with a sufficiently severe sanction, rational would-be criminals desist from crime. By the very fact that crime no longer pays, the actual enforcement of the sanction becomes an action off the equilibrium path. The mere threat is enough. From this perspective, imperfect sanctions are not only pointless, they are even counter-productive. Rational criminals realize that crime is profitable business. If society does not want to entirely give up on its normative expectations, a lot of costly enforcement becomes necessary precisely because the sanction is not perfect.

This paper tries to rebut this argument on its own turf, i.e. with an argument from rational choice theory. I introduce sanctions into a standard model of social preferences. With the help of this very simple model I show that imperfect sanctions do serve a socially desirable purpose if only a sufficient fraction of the population is not completely selfish but also cares about the negative effects of their own actions on the well-being of others. The theoretical argument even goes through if the population is known to be heterogeneous, with some selfish individuals. The theoretical argument is also robust to the introduction of mild uncertainty about the willingness of others to forego additional benefit for themselves at the expense of others.

I use a standard experimental design, the linear public good with punishment, to test two predictions resulting from my theory: if participants hold social preferences, imperfect sanctions

help them sustain cooperation. This even holds if group members face complete freeriders. Both predictions are supported by the data.

It has been the purpose of my paper to test the theoretical expectation that imperfect sanctions are instrumental if sufficiently many individuals hold social preferences. Yet my results also qualify the finding from Tyran and Feld (2006) that exogenously imposed imperfect sanctions are pointless. My experiment differs in a number of respects. All of them might, either independently or jointly, explain why I do find the effect. In my case, the sanction is exogenous, but meted out by a human subject. Consequently there is strategic uncertainty about the punishment policy of the authority currently in charge. The authority has to pay for punishment. I test contributions in a repeated game. The exogenously induced norm is not maximum contributions, but a more realistic measure, taken from an earlier experiment in the same lab. Moreover, my empirical strategy differs. My dependent variable is not contribution choices in the expectation of a punishment scheme, but changes in contributions in reaction to the experience of punishment. Later experiments might want to systematically vary these and further factors, in the interest of defining under which conditions imperfect sanctions are still effective, and how marginal reactions to severity interact with these situational features.

Lab experiments are not meant to map reality. Exploiting other anonymous group members in a public good game is not crime. There is not even an explicit normative expectation (although informing participants about results from an earlier experiment with a very similar setup was meant to induce an implicit normative expectation). On purpose economic experiments are unframed. The naked opportunity structure does not trigger moral compunctions the same way as those real life situations that trigger criminal investigations or tort liability. In the lab, no more than a few cents are at stake, whereas the effects of crime or tort on victims tend to be much more severe. Gains from misbehavior are also much more contained in the lab than they tend to be in the field. Therefore socially desirable behavior might be cheaper in the lab.

I am happy to acknowledge all these limitations inherent in my empirical approach. In the light of these limitations, one certainly should be hesitant to directly extrapolate from these findings to policy choices. All I hope to contribute to the policy discourse is one argument. If it is reasonable to expect heterogeneous preferences and a substantial fraction of the relevant population not being straightforwardly selfish, then imperfect sanctions may serve a useful purpose. They deter those who are somewhat inclined to care about harm on others, but not strongly enough. The prospect of mild sanctions may suffice to tilt the balance in favor of socially desirable behavior. Slightly more severe, but still imperfect sanctions may suffice if the willingness of others to play by the rules is uncertain, or if a few are known to misbehave. Mild sanctions are of course not enough to deter the latter. But they may help prevent occasional acts of antisocial behavior to trigger a vicious cycle. Of course, society is even better off if it inflicts perfect sanctions on those who do not care about others in the first place. But this requires discriminating according to social preferences, which may be difficult to do in the field. In that event, mild sanctions may be an affordable way of making the majority of society resilient against the occasional experience of deviance.

References

AMBRUS, ATTILA and BEN GREINER (2011). Imperfect Public Monitoring with Costly Punishment. An Experimental Study http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1959170.

ANDREONI, JAMES and B. DOUGLAS BERNHEIM (2009). "Social Image and the 50-50 Norm. A Theoretical and Experimental Analysis of Audience Effects." Econometrica **77**: 1607-1636.

ARNEKLEV, BRUCE J., HAROLD G. GRASMICK, CHARLES R. TITTLE and ROBERT J. BURSIK (1993). "Low Self-control and Imprudent Behavior." Journal of Quantitative Criminology **9**(3): 225-247.

BALLIET, DANIEL, LAETITIA B MULDER and PAUL AM VAN LANGE (2011). "Reward, Punishment, and Cooperation: A Meta-analysis." Psychological Bulletin **137**(4): 594-615.

BLANCO, MARIANA, DIRK ENGELMANN and HANS-THEO NORMANN (2011). "A Within-Subject Analysis of Other-Regarding Preferences." Games and Economic Behavior **72**: 321-338.

BOLTON, GARY E. and AXEL OCKENFELS (2000). "ERC: A Theory of Equity, Reciprocity and Competition." American Economic Review **90**: 166-193.

CARPENTER, JEFFREY P (2007). "The Demand for Punishment." Journal of Economic Behavior & Organization **62**(4): 522-542.

CARPENTER, JEFFREY P, PETER HANS MATTHEWS and OKOMBOLI ONG'ONG'A (2004). "Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms." Journal of Evolutionary Economics **14**(4): 407-429.

CASARI, MARCO (2005). "On the Design of Peer Punishment Experiments." Experimental Economics **8**(2): 107-115.

CHARNESS, GARY (2000). "Self-Serving Cheap Talk. A Test of Aumann's Conjecture." Games and Economic Behavior **33**: 177-194.

CHARNESS, GARY and MATTHEW RABIN (2002). "Understanding Social Preferences with Simple Tests." Quarterly Journal of Economics **117**: 817-869.

CHAUDHURI, ANANISH (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments. A Selective Survey of the Literature." Experimental Economics **14**: 47-83.

COCHRAN, JOHN K., PETER B. WOOD and BRUCE J. ARNEKLEV (1994). "Is the Religiosity-Delinquency Relationship Spurious? A Test of Arousal and Social Control Theories." Journal of Research in Crime and Delinquency **31**(1): 92-123.

CROSETTO, PAOLO, WERNER GÜTH, LUIGI MITTONE and MATTEO PLONER (2012). Motives of Sanctioning: Equity and Emotions in a Public Good Experiment with Punishment http://pubdb.wiwi.uni-jena.de/pdf/wp_2012_046.pdf.

DE LI, SPENCER (2004). "The Impacts of Self-control and Social Bonds on Juvenile Delinquency in a National Sample of Midadolescents." Deviant Behavior **25**: 351-373.

DUFWENBERG, MARTIN, SIMON GÄCHTER and HEIKE HENNIG-SCHMIDT (2011). "The Framing of Games and the Psychology of Play." Games and Economic Behavior **73**(2): 459-478.

DUFWENBERG, MARTIN and GEORG KIRCHSTEIGER (2004). "A Theory of Sequential Reciprocity." Games and Economic Behavior **47**: 268-298.

EGAS, MARTIJN and ARNO RIEDL (2008). "The Economics of Altruistic Punishment and the Maintenance of Cooperation." Proceedings of the Royal Society B: Biological Sciences **275**(1637): 871-878.

ENGEL, CHRISTOPH and BERND IRLENBUSCH (2010). Turning the Lab into Jeremy Bentham's Panopticon. The Effect of Punishment on Offenders and Non-Offenders http://ssrn.com/abstract=1555589.

ENGEL, CHRISTOPH and LILIA ZHURAKHOVSKA (2012). You are in Charge. Experimentally Testing the Motivating Power of Holding a (Judicial) Office

ENGELMANN, DIRK and MARTIN STROBEL (2004). "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." American Economic Review **94**: 857-869.

ESBENSEN, FINN-AAGE and ELIZABETH PIPER DESCHENES (1998). "A Multisite Examination of Youth Gang Membership. Does Gender Matter?" Criminology **36**: 799-828.

FALK, ARMIN and URS FISCHBACHER (2006). "A Theory of Reciprocity." Games and Economic Behavior **54**: 293-315.

FEHR, ERNST and SIMON GÄCHTER (2000). "Cooperation and Punishment in Public Goods Experiments." American Economic Review **90**: 980-994.

FEHR, ERNST and KLAUS M. SCHMIDT (1999). "A Theory of Fairness, Competition, and Cooperation." Quarterly Journal of Economics **114**: 817-868.

FISCHBACHER, URS (2007). "z-Tree. Zurich Toolbox for Ready-made Economic Experiments." Experimental Economics **10**: 171-178.

GLAZE, LAUREN E. and THOMAS P. BONCZAR (2007). Probation and Parole in the United States, 2006 http://www.ojp.usdoj.gov/bjs/pub/pdf/ppus06.pdf.

GRECHENIG, KRISTOFFEL, ANDREAS NICKLISCH and CHRISTIAN THÖNI (2010). "Punishment Despite Reasonable Doubt. A Public Goods Experiment with Sanctions Under Uncertainty." Journal of Empirical Legal Studies **7**(4): 847-867.

GREINER, BEN (2004). An Online Recruiting System for Economic Experiments. Forschung und wissenschaftliches Rechnen 2003. K. Kremer and V. Macho. Göttingen**:** 79-93.

GÜRERK, ÖZGÜR, BETTINA ROCKENBACH and IRENAEUS WOLFF (2010). The Effects of Punishment in Dynamic Public-good Games http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1589362.

HECKATHORN, DOUGLAS D. (1989). "Collective Action and the Second-Order Free-Rider Problem." Rationality and Society **1**: 78-100.

HOLT, CHARLES A. and SUSAN K. LAURY (2002). "Risk Aversion and Incentive Effects." American Economic Review **92**: 1644-1655.

JOHNSON, TIM, CHRISTOPHER T DAWES, JAMES H FOWLER, RICHARD MCELREATH and OLEG SMIRNOV (2009). "The Role of Egalitarian Motives in Altruistic Punishment." Economics Letters **102**(3): 192-194.

KERLEY, KENT R., XHIAOHE XU and BANGON SIRISUNYALUCK (2008). "Self-control, Intimate Partner Abuse, and Intimate Partner Victimization. Testing the General Theory of Crime in Thailand." Deviant Behavior **29**: 503-532.

KOSFELD, MICHAEL, AKIRA OKADA and ARNO RIEDL (2009). "Institution Formation in Public Goods Games." American Economic Review **99**: 1335-1355.

LAGRANGE, TERESA C. and ROBERT A. SILVERMAN (1999). "Low Self-Control and Opportunity. Testing the General Theory of Crime as an Explanation for Gender Differences in Delinquency." Criminology **37**: 41-72.

LEDYARD, JOHN O. (1995). Public Goods. A Survey of Experimental Research. The Handbook of Experimental Economics. J. H. Kagel and A. E. Roth. Princeton, NJ, Princeton University Press**:** 111-194.

LEVITT, STEVEN D. and SUDHIR ALLADI VENKATESH (2007). An Empirical Analysis of Street-Level Prostitution http://economics.uchicago.edu/pdf/Prostitution%205.pdf?q=venkatesh.

LIEBRAND, WIM B. and CHARLESG. MCCLINTOCK (1988). "The Ring Measure of Social Values. A Computerized Procedure for Assessing Individual Differences in Information Processing and Social Value Orientation." European Journal of Personality **2**: 217-230.

MASCLET, DAVID, CHARLES NOUSSAIR, STEVEN TUCKER and MARIE-CLAIRE VILLEVAL (2003). "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism." American Economic Review **93**: 366-380.

MASCLET, DAVID and MARIE-CLAIRE VILLEVAL (2008). "Punishment, Inequality, and Welfare. A Public Good Experiment." Social Choice and Welfare **31**(3): 475-502.

MONTERO, MARIA, MARTIN SEFTON and PING ZHANG (2008). "Enlargement and the Balance of Power. An Experimental Study." Social Choice & Welfare **30**: 69-87.

NIKIFORAKIS, NIKOS S. and HANS-THEO NORMANN (2008). "A Comparative Statics Analysis of Punishment in Public Good Experiments." Experimental Economics **11**: 358-369.

RABIN, MATTHEW (1993). "Incorporating Fairness into Game Theory and Economics." American Economic Review **83**: 1281-1302.

SCHWARTZ, GARY (1994). "Reality in the Economics of Tort Law. Does Tort Law Really Deter?" UCLA Law Review **42**: 377-444.

SOUSA, SERGIO (2010). Cooperation and Punishment under Uncertain Enforcement http://ideas.repec.org/p/cdx/dpaper/2010-06.html.

THÖNI, CHRISTIAN (2011). Inequality Aversion and Antisocial Punishment, Department of Economics, University of St. Gallen.

TYRAN, JEAN-ROBERT and LARS P. FELD (2006). "Achieving Compliance when Legal Sanctions are Non-Deterrent." Scandinavian Journal of Economics **108**: 135-156.

WILSON, JAMES Q. and ALLAN ABRAHAMSE (1992). "Does Crime Pay?" Justice Quarterly **9**: 359-377.

WINFREE, L. THOMAS and FRANCES P. BERNAT (1998). "Social Learning, Self-control, and Substance Abuse by Eighth Grade Students. A Tale of Two Cities." Journal of Drug Issues **28**: 539-558.

YAMAGISHI, TOSHIO (1986). "The Provision of a Sanctioning System as a Public Good." Journal of Personality and Social Psychology **51**: 110-116.

ZELMER, JENNIFER (2003). "Linear Public Goods. A Meta-Analysis." Experimental Economics **6**: 299-310.

Appendix

Prediction of Authority Choices

In the companion paper we show that authorities feel under an obligation to manage the groups to which they happen to be assigned (Engel and Zhurakhovska 2012). One way of formalizing this is an efficiency motive, as in (Charness and Rabin 2002**:** 852). The easiest way of writing authorities' payoff function is

$\pi_{at} = e_a - c\sum_i s_i$	(12)

where a stands for the authority in question, e_a is the authority's gross endowment, i.e. including the additional payoff in case she does not use any punishment points, s_i is the punishment points meted out to active participant i, and c is the cost of any punishment point, for the authority. Obviously, $s_i^* = 0$. The authority keeps all her punishment points.

Now assume that the authority not only cares about her own payoff, but also about efficiency, so that we have

$u_{at} = (1-\vartheta)\pi_{at} + \vartheta(\pi_{at} + \sum_i \pi_{it})$	(13)

with parameter ϑ expressing how much weight she attaches to efficiency. From (1), the efficient contribution choice of active players is $c_{it} = e$. If the authority expects or learns active players' aversion against advantageous inequity to be so pronounced that all will contribute their entire endowments, even if there is no punishment, irrespective of the size of punishment, the authority will not punish at all. Note that this choice is efficient. Welfare increases for two reasons: the authority saves the cost of punishment and, more importantly, the active players in question save the loss from punishment. However if the authority is more pessimistic, or if she learns that there are free-riders, and if she attaches enough weight to efficiency, i.e. if ϑ is large enough, then she will mete out punishment such that active players contribute fully. What this means depends on the expected or observed preferences of active players. If they are inequity averse, an efficiency minded authority will not mete out maximum punishment, but will content herself with imperfect sanctions. On two channels this is more efficient: the authority saves money, and active players suffer less from punishment. This yields

P_4: a) If the authority derives sufficient utility from efficiency, she punishes the active players who do not make full contributions.
b) If the authority expects or learns active players to be averse against advantageous inequity, she metes out imperfect sanctions.

Instructions

General Instructions

In the following experiment, you can earn a substantial amount of money, depending on your decisions. It is therefore very important that you read these instructions carefully.

During the experiment, any communication whatsoever is forbidden. If you have any questions, please ask us. Disobeying this rule will lead to exclusion from the experiment and from all payments.

You will in any case receive 4 € for taking part in this experiment. In the first two parts of the experiment, we do not speak not of €, but instead of Taler. Your entire income from these two parts of the experiment is hence initially calculated in Taler. The total number of Taler you earn during the experiment is converted into € at the end and paid to you in cash, at the rate of

1 Taler = 4 Eurocent.

The experiment consists of four parts. We will start by explaining the first part. You will receive separate instructions for the other parts.

Part One of the Experiment

In the first part of the experiment, there are two roles: A and B. Four participants who have the role A form a group. One participant who has the role B is allocated to each group. The computer will randomly assign your role to you at the beginning of the experiment.

On the following pages, we will describe to you the exact procedure of this part of the experiment.

Information on the Exact Procedure of the Experiment

This part of the experiment has two steps. In the first step, role A participants make a decision on contributions to a project. In the second step, the role B participant can reduce the role A participants' income. At the start, each **role A** participant receives **20 Taler**, which we refer to in the following as the **endowment**. **Role B** participants receive 20 points at the start of step 2. We explain below how role B participants may use these points.

Step 1:

In Step 1, **only the four role A participants** in a group make a decision. Each role A member's decision influences the income of all other role A players in the group. The income of player B is not affected by this decision. As a role A participant, you have to decide how many of the 20 Taler you wish to invest in a **project** and how many you wish to keep for yourself.

If you are a **role A** player, **your income** consists of two parts:

(1) the Taler you have kept for yourself (**"income retained from endowment")**
(2) the **"income from the project"**. The income from the project is calculated as follows:

Your income from the project =
0.4 *times* the total sum of contributions to the project

Your **income** is therefore calculated as follows:

(20 Taler – your contribution to the project) + 0.4* (total sum of contributions to the project).

The income **from the project** of all role A group members is calculated according to the same formula, i.e., each role A group member receives the same income from the project. If, for example, the sum of the contributions from all role A group members is 60 Taler, then you and all other role A group members receive an income from the project of 0.4*60 = 24 Taler. If the role A group members have contributed a total of 9 Taler to the project, then you and all other role A group members receive an income from the project of 0.4*9 = 3.6 Taler.

For every Taler that you keep for yourself, you earn an income of 1 Taler. If instead you contribute a Taler from your endowment to your group's project, the sum of the contributions to the project increases by 1 Taler and your income from the project increases by 0.4*1 = 0.4 Taler. However, this also means that the income of all other role A group members increases by 0.4 Taler, so that the total group income increases by 0.4*4 = 1.6 Taler. In other words, the other role A group members also profit from your own contributions to the project. In turn, you also benefit from the other group members' contributions to the project. For every Taler that another group member contributes to the project, you earn 0.4*1 = 0.4 Taler.

Please note that the role B participant cannot contribute to the project and does not earn any income from the project.

Step 2:

In Step 2, **only the role B participant** makes decisions. As role B participant, you may **reduce or maintain** the income of **every** participant in Step 2 by distributing **points**.

At the beginning of Step 2, the four role A participants and the role B participant are told how much each of the role A participants has contributed to the project.

As a role B player, you now have to decide, for **each** of the four role A participants, whether you wish to distribute points to them and, if so, how many points you wish to distribute to them. You are obliged to enter a figure. If you do not wish to change the income of a particular role A participant, please enter 0. Should you choose a number greater than zero, you reduce the income of that particular participant. **For each point that you allocate to a participant, the income of this participant is reduced by 3 Taler.**

The total Taler income of a role A participant from both steps is hence calculated using the following formula:

Income from Step 1 – 3* (sum of *points* received)

Please note that Taler income at the end of Step 2 can also be negative for role A participants. This can be the case if the income-subtraction from points received is larger than the income from Step 1. However, the role B participant can distribute a maximum of 20 points to all four role A members of the group. 20 points are the maximum limit. As a role B participant, you can also distribute fewer points. It is also possible not to distribute any points at all.

If you have role B, please state your reasons for your decision to distribute (or not to distribute) points, and why you distributed a particular number of points, if applicable. In doing this, please try to be factual. Please enter your statement in the corresponding space on your screen. You have 500 characters max. to do this. Please note that, in order to send your statement, you will have to press "Enter" once each time. As soon as you have done this, you will no longer be able to change what you have written.

The income of the role B participant does not depend on the income of the other role A participants, nor on the income from the project. For taking part in the first part of the experiment, he or she receives a fixed payment of

1 €.

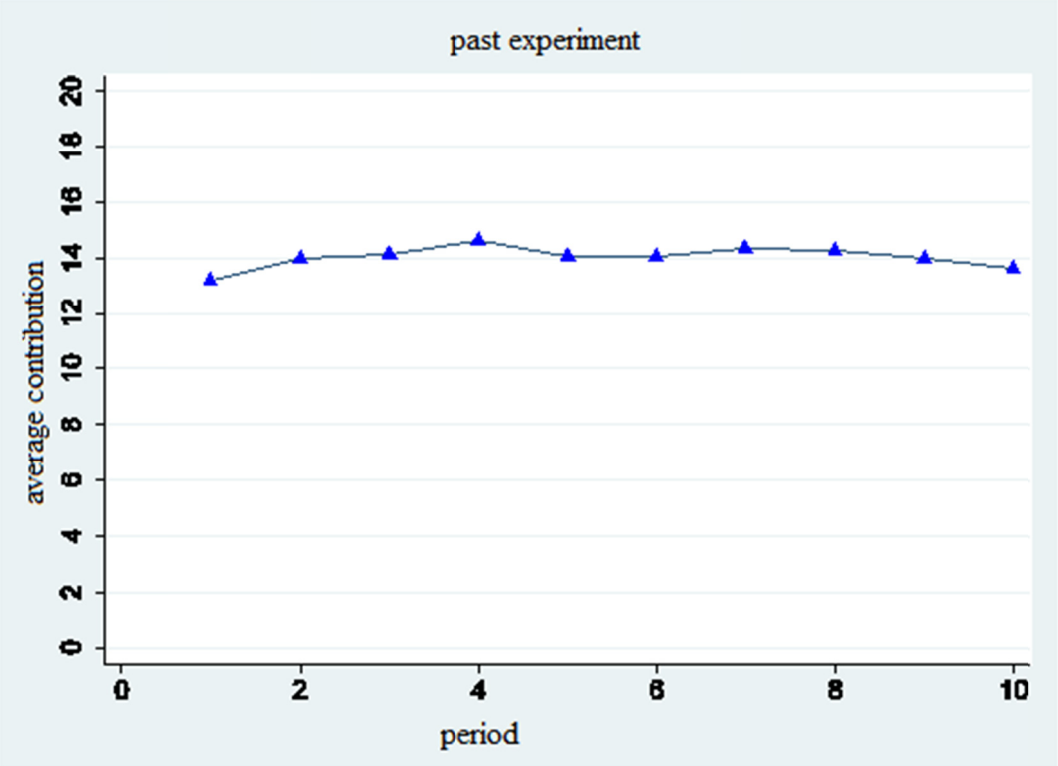
In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.

After this, the first part of the experiment ends. You will then be told what your payment is for this part of the experiment. Hence, you will also know how many points you and all other participants have been given by player B.

Experiences from an Earlier Experiment

For your information, we give you the following graph, which tells you the average contributions made in a very similar experiment that was conducted in this laboratory.

In this experiment, too, there were groups of 4 role A participants and one role B participant each. The role A participants' income was calculated in exactly the same way. The experiment had 10 equal periods. The role B participant also had 20 points at his disposal in each period. At the end of each period, the role A participants were told how much each of the other participants had contributed and how the role B participant had reacted to this.



Part Two of the Experiment

The second part of the experiment consists of 10 repetitions of the first part. **Throughout the entire second part, all participants keep the role they had in the first part of the experiment. The computer randomly rematches the groups of four in every period. In each period, the computer randomly assigns a role B participant to each group.**

As a reminder:
In each period, each role A participant receives 20 Taler, which may be contributed to the project entirely, in part, or not at all. For each period, calculating the income from the project for the role A participants in a group happens in exactly the same way as it did in the first part of the experiment. In each period, each role B participant receives 20 points, which may be used to reduce the income of the players A in the group. For each point that a role A participant receives in a period, 3 Taler are subtracted. For each point that a role B participant does not use, he or she is given the sum of 0.01 €. In addition to the income from the points retained, each role B participant receives a flat fee of 10 € for participating in this second part of the experiment.

At the beginning of Step 2 of each period, the four role A participants and the role B participant are told how much each of the role A participants contributed to the project.

Please note that the groups are rematched anew in each period.

After each period, you are told about your individual payoff. You are therefore also informed how many points you and the other participants have been assigned by the role B participant.

Part Three of the Experiment

We will now ask you to make some decisions. In order to do this, **you will be randomly paired with another participant.** In several distribution decisions, you will be able to allocate points to this other participant and to yourself by repeatedly **choosing between two distributions, 'A' and 'B'**. The points you allocate to yourself will be paid out to you at the end of the experiment at a rate of **500 points = 1 €**. At the same time, you are also randomly assigned to **another** participant in the experiment, who is, in turn, also able to allocate points to you by choosing between distributions. This participant is **not the same participant** as the one to whom you have been allocating points. The points allocated to you are also credited to your account. The **sum** of all points you have allocated to yourself and those allocated to you by the other participant are paid out to you at the end of the experiment at a rate of 500 points = 1 €.

Please note that the participants assigned to you in this part of the experiment are **not the members of your group** from the preceding part of the experiment. You will therefore be dealing with other participants.

The individual decision tasks will look like this:

Possibility A:		**Possibility B:**	
Your points	The points of the experiment participant allocated to you	Your points	The points of the experiment participant allocated to you
0	500	304	397

A B

In this example: If you click 'A', you give yourself 0 points and 500 points to the participant allocated to you. If you click 'B', you give yourself 304 points and 397 points to the participant allocated to you.