Fostering cooperation through the enhancement of own vulnerability

Kopányi-Peuker, A.; Offerman, T.; Sloof, R.

Fostering cooperation through the enhancement of own vulnerability

Anita Kopányi-Peuker, Theo Offerman and Randolph Sloof*

October 4, 2013

Abstract

We consider the possibility that cooperation in a prisoner's dilemma is fostered by people's voluntary enhancement of their own vulnerability. The vulnerability of a player determines the effectiveness of possible punishment by the other. In the "Gradual" mechanism, players may condition their incremental enhancements of their vulnerability on the other's choices. In the "Leap" mechanism, they unconditionally choose their vulnerability. In our experiment, subjects only learn to cooperate when either one of these mechanisms is allowed. In agreement with theory, subjects aiming for cooperation choose higher vulnerability levels in Gradual than in Leap, which maps into higher mutual cooperation levels.

JEL classification: D03, D81, D83

Keywords: prisoner's dilemma, cooperation, endogenous punishment.

Acknowledgements: This paper benefited from the suggestions of audiences at the University of Zurich, University of California Santa Barbara, Friedrich Schiller University Jena, the University of Paris 1, the ESA European Conference 2012 (University of Cologne), the 5$^{\text{th}}$ Maastricht Behavioral and Experimental Economics Symposium (Maastricht University), the ACLE Competition & Regulation Meeting 2012 (University of Amsterdam) and 2011 Workshop on Trust and Cultural Evolution (University of Valencia and ERI-CES). Financial help of the Research Priority Area Behavioral Economics of the University of Amsterdam is gratefully acknowledged.

*CREED, University of Amsterdam and Tinbergen Institute, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. We are very grateful to CREED programmer Jos Theelen for programming the experiment. A.G.Kopanyi-Peuker@uva.nl; T.J.S.Offerman@uva.nl; R.Sloof@uva.nl

1 Introduction

In situations where private interests are at odds with the collective interest, the possibility to punish has proven to be an effective tool to support cooperation. In small groups, contributors are often willing to pay a small cost to punish free riders. This process helps to make free riders behave in agreement with the collective interest (see for instance Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2000; 2002). In practice, the possibility to punish free riders will often be limited though, because people are protected by property rights. For instance, a neighbor who refrains from contributing to the local public good can presumably be disciplined if his car is damaged, but the person who inflicts the damage has to face the possibility that she will be persecuted in court.

Here we consider a prisoner's dilemma in which cooperation cannot be contracted and in which the possibility to punish may only endogenously become available. That is, players can only be punished if they voluntarily make themselves vulnerable in the first place. They may do so to signal that they are interested in pursuing mutual cooperation. If the other is willing to punish a free rider at a small cost, then the player's signal to conditionally cooperate becomes credible. Theoretically and in an experiment, we investigate whether and how the possibility of enhancing one's own vulnerability may foster cooperation in the prisoner's dilemma. In particular, we want to know whether (i) people voluntarily make themselves vulnerable if they have the possibility to do so, (ii) whether cooperation is enhanced when players have made themselves vulnerable and (iii) whether it matters if the trust-building process occurs gradually or in one single step.

The trust-building process that we have in mind corresponds to how strangers are reported to build friendships. In his "Moralia", Plutarch already described how a reciprocal exchange of secrets may lead to a relationship in which there is a fear of loss of trust (Gambetta, 2009, p. 66; Plutarch, 1992). Strangers who have exchanged secrets before they interact in a prisoner's dilemma may refrain from free riding if they fear that this will trigger the other to publicly disclose the secret. Recent psychological research shows that feelings of intimacy develop in a dynamic process in which a person discloses personal information, thoughts and secrets and the partner responds in a likewise manner (Altman, 1973; Rotenberg, 1986; Dindia and Allen, 1992; Laurenceau et al., 1998). An important (but possibly unintended) by-product of such feelings may be that people have learned to trust each other when they subsequently interact in a prisoner's dilemma.[1] Interestingly, Derlega et al. (1976) report that reciprocal self-disclosure is especially observed among strangers, as hypothesized by Altman (1973). Once relationships have been established and friends have learned

[1] Gossip may also serve as a form of self-disclosure that promotes trust through the enhancement of vulnerability. If a worker communicates damaging information about a superior to a colleague, he faces the risk that the colleague reveals this to the superior. If instead the colleague reciprocates with another negative story, a bond may be formed which may help the workers to solve free rider incentives in the work place. See Sommerfeld et al. (2007) for a discussion of other functions of gossip.

to trust each other, other arguments may take over. For instance, in long term relations the behavior of friends may be disciplined by repeated game considerations.[2]

In this paper, we model this trust-building process in a stylized three stage game. In the first stage, the two players voluntarily decide upon the extent to which they make themselves vulnerable for punishment. In the second stage, after having been informed of each other's own possible punishment level, the two partners decide whether or not to cooperate in a prisoner's dilemma. Based on the observed outcome, each player decides in the third stage whether or not to punish the partner at a small cost. If a player decides to punish, the partner loses an amount equal to her chosen own possible punishment level in the first stage.

We consider two different trust-building mechanisms. In the "Gradual" variant, the players may build trust in small steps, while observing the partner's willingness to go along in this process. This variant has the advantage that a player can condition the own possible punishment level on the partner's possible punishment level. It agrees with the empirical observation that trust is often formed in small incremental steps. The possibility of "starting small" can be advantageous, as has been shown by Andreoni and Samuelson (2006) for a repeated prisoner's dilemma and by Weber (2006) in a team production game with Pareto ranked equilibria.[3]

In the "Leap" variant, players decide whether or not to take a leap in the dark by simultaneously choosing an own possible punishment level without the possibility of conditioning it on the partner's level. This variant may correspond to situations where players do not have the time to build trust in small steps, or where it is too costly to engage in a slow gradual process.

We first provide a theoretical analysis of the two trust-building mechanisms. We show that in either mechanism mutual defection as well as mutual cooperation can be supported in equilibrium. The cooperative equilibrium is selected in a process of iterated elimination of weakly dominated strategies in the Gradual mechanism. Such a process does not have a bite in the Leap mechanism

[2]Other applications of the mechanisms studied in this paper include the use of hostages and trust-building among criminals. Since Roman times, hostages have been exchanged to enforce truces and treaties. In some cases hostages were voluntarily exchanged (Schelling, 1960, p. 135-137; Lee, 1991; Herrmann and Palmieri, 2005). Williamson (1983) discusses how the exchange of hostages can be used to support trade in contractual relations hampered by holdup threats; he argues that "... the use of hostages to support exchange is widespread and economically important" (p. 537). Gambetta (2009) describes some examples where trust-building mechanisms are used to support criminal activities. For instance, mafia bosses have been reported to bring their wives to potentially explosive dinners to signal their own willingness not to start a shooting. Pedophiles are often asked to share compromising photos before they get access to a child-pornography website.

[3]The mechanism behind the result in those papers is quite different though. In Andreoni and Samuelson (2006), players differ in their taste for cooperation when they participate in a twice-played prisoner's dilemma. If the stakes are larger in the second stage game, players have a larger willingness to invest in the first stage to achieve cooperation in the second. In Weber (2006), coordination on the efficient equilibrium is facilitated if teams start small and new members observe the history of the team before they enter.

though. Further, when weakly dominated strategies are eliminated iteratively, players choose higher own punishment levels in the Gradual variant than they would ever do in the Leap variant. In essence, the possibility to condition one's own vulnerability on the partner's vulnerability allows players to turn themselves into unconditional cooperators in the Gradual variant, provided that the partner does the same. Theoretically, such high levels of trust cannot be reached in the Leap variant.

We test the performance of the mechanisms in an experiment in which we use a harsh strangers environment. In a Control treatment where players do not have the possibility to make themselves vulnerable, cooperation is not sustained. In the treatments with a trust-building mechanism, after some initial aversion a significant fraction of the subjects actively employs the mechanism to achieve cooperative outcomes. The two trust-building treatments do not differ in the extent to which subjects use the mechanism. Instead, conditional on the mechanism being used, subjects choose higher own possible punishment levels in the Gradual variant than in the Leap variant. These higher own possible punishment levels subsequently map into a higher frequency of cooperative outcomes in the Gradual variant. At the same time, by helping to align the players' vulnerability levels, the Gradual variant diminishes the occurrence of miscoordination outcomes in which one player cooperates and the other defects.[4]

In the second half of the experiment, cooperative outcomes are regularly observed with a mechanism but almost never in the Control treatment. After subjects have gained experience with the environment and the mechanism, subjects cooperate in only 4.3% of the cases in the Control treatment, while they cooperate in 26.5% of the cases in the Leap variant and in 36.3% of the cases in the Gradual variant. The Gradual mechanism performs well, even when it is put in the perspective of studies in which the punishment possibility was introduced exogenously. For instance, in Fehr and Gächter's (2000) strangers treatment, subjects' contributions increased from 18.5% to 57.5% when subjects were exogenously allowed to punish. So in absolute terms the effect of introducing the possibility to punish is somewhat bigger in Fehr and Gächter (2000), but not in relative terms.

The remainder of the paper is organized in the following way. Section 2 discusses the related literature. Section 3 presents the game and the theoretical analysis. In Section 4, we provide the experimental design and procedures. Section 5 presents the experimental results and Section 6 concludes.

[4]Our results complement the findings of Potters and Suetens (2009) who find more cooperation when actions exhibit strategic complementarities than when they display strategic substitutes. Both our Leap and Gradual mechanisms are examples of interactions with strategic complementarities. Our paper shows that within the class of games with strategic complementarities, cooperative behavior may result more easily if players have the possibility to reciprocate in small steps.

2 Related Literature

Our study contributes to several strands of literature. One related literature considers how agents may achieve cooperation if they can write binding contracts. In the spirit of the work of Coase (1960), Varian (1994) proposes a simple two-stage compensation scheme that implements the efficient outcome as a subgame perfect equilibrium in a wide class of games. In a prisoner's dilemma, the mechanism amounts to the following. In the first stage, each player announces a non-negative amount that he will pay to the partner provided that the partner cooperates in the second stage. In the second stage, the players participate in a prisoner's dilemma. An announcement made in the first stage is binding, so if the partner cooperates the contract is automatically carried out.

In an experiment, Andreoni and Varian (1999) show that the compensation mechanism performs fairly well. When the mechanism is introduced in the second part of the experiment, cooperation levels go up from 25.8% to 50.5%. So the data provide support for the mechanism, even though the mechanism does not completely weed out defective choices.

Charness et al. (2007) provide a more demanding test of the compensation mechanism. They extend Andreoni and Varian's experimental results in two ways. First, they consider variants of the prisoner's dilemma in which there is a substantial range of transfers that support cooperation in a subgame perfect equilibrium (in the prisoner's dilemma of Andreoni and Varian, there was essentially a unique subgame perfect equilibrium). Second, in their variants of the prisoner's dilemma subjects face a coordination game in the second stage of the experiment; both (C,C) and (D,D) can be supported as equilibria of the second subgame (in contrast, in Andreoni and Varian subjects have a dominant strategy to contribute provided that sufficient transfers in the first stage are made). The compensation mechanism even performs well in this harsher environment. Charness et al. find cooperation rates in the range of 43-68% when transfer payments are permitted compared to 11-18% in the control without the compensation scheme.

In a more complicated emissions trading game, Hamaguchi et al. (2003) investigate a punishment version of Varian's compensation scheme and find less support for the mechanism. In their game, there are many Nash equilibria, which may have inhibited coordination on the subgame perfect equilibrium. Chen and Gazzale (2004) show that supermodularity can be an important factor in the performance of a generalized version of the compensation scheme; they find that supermodular games converge significantly better than games far below the threshold of supermodularity. Falkinger (1996) introduces a related mechanism in which players are subsidized or taxed in accordance with how their contribution to the public good deviates from the mean contribution. In an experiment, Falkinger et al. (2000) confirm the potential of the mechanism to support cooperation in public goods.

Probably closest to us in terms of mechanism are Iossa and Spagnolo (2011). In a theoretical contribution, they discuss a contractual clause that may help players to achieve cooperation in a non-verifiable task. Prior to the interaction,

the players sign a contract that allows them to punish each other to a certain extent at their own discretion. The punishment is without value to the punisher but harms the partner at whom it is aimed.[5] The main difference between their and our approach lies in how the first stage is modeled. In their paper, players have the possibility to write a binding contract in the first stage, so that they can agree on a mutually beneficial arrangement without strategic uncertainty. In contrast, in our non-cooperative approach, players have to deal with the danger that their choice to enhance the own vulnerability is not matched by the partner. Among other things, this implies that in our approach non-use of the mechanism in combination with mutual defection cannot be excluded in equilibrium.

We contribute to this literature in two ways. First, an attractive feature of the mechanism that we study is that it does not require the possibility to write binding contracts. Second, we show that it may matter a lot for the outcome if players have the possibility to build trust in many small steps. Our conjecture is that this finding will generalize to Varian's compensation scheme, and that higher levels of cooperation may be achieved if the first contract writing stage proceeds gradually.

Our paper also contributes to a somewhat more remote strand of literature that studies how subjects choose among existing institutions that do or do not allow for punishment. In the public good game of Gürerk et al. (2006), subjects can choose between a sanction-free institution and a sanctioning institution in which punishments and rewards are allowed. After an initial phase in which a small majority prefers the community without sanctioning, subjects massively migrate to the community in which sanctions are allowed and where high levels of cooperation are sustained throughout the experiment. In a follow-up experiment, Gürerk et al. (2009) show that the sanctioning institution remains equally successful if punishments but no rewards are allowed, and they find that initial self-selection in the sanctioning institution is a key-factor explaining its success.

Kosfeld et al. (2009) investigate a social dilemma in which sanctions are imposed by a central authority. In their three stage game, players first decide whether or not to participate in the organization. Non-members of the organization can free ride on the members of the organization. In the second stage, players are informed of the number of participants in the organization and the organization is only actually formed if all members agree. In the third stage, all players decide about their contribution decision and members of the organization are automatically punished if they do not contribute their full endowment

[5] The contractual clause studied by Iossa and Spagnolo (2011) thus has the defining characteristics of a hostage (cf. Schelling, 1960). Contractual hostages to facilitate trade were first studied theoretically in the economics literature by Williamson (1983). Inspired by his work, a small literature emerged in sociology that investigates so-called "hostage games". In essence, the hostage here is a contract that specifies a punishment that is automatically administered to the poster of the hostage if he acts opportunistically in the subsequent social dilemma. Raub and Keren (1993) show that posting of such hostages in a stage prior to a prisoners' dilemma may enhance cooperation. Likewise, Snijders and Buskens (2001) find that the postage of hostages may also help achieving cooperation in the trust game. Raub (2009) surveys the relevant sociological literature.

to the public good. In the experiment, the grand organization is usually formed in which everybody participates. This result is in line with the notion that social preferences matter and is inconsistent with standard theory that predicts the formation of incomplete organizations.

In the public good game studied by Sutter et al. (2010), subjects vote to play in a standard Voluntary Contribution Mechanism (VCM), a VCM that allows for rewards and a VCM that allows for punishments. When rewards and punishments are sufficiently effective, subjects prefer the VCM with rewards, even though the VCM with punishments is more effective. Sutter et al. find that endogenously chosen institutions trigger higher cooperation levels than if the same institution is implemented exogenously.

These studies share the feature that once an institution is voluntarily implemented, its effectiveness is exogenously determined (by the experimenter).[6] In contrast, in our study subjects themselves determine how effective a punishment by the partner will be.[7]

3 The model and its predictions

3.1 Structure of the game

In this section we present a simple model on which our experiment is based. The model consists of a two-player game with three stages. In the first stage players unilaterally determine the level of possible punishment they might receive themselves later on in the game. We consider two different ways in which these possible punishments can be chosen. In the *Gradual mechanism*, players' own possible punishment levels are gradually increased by an automatic process as in an (Japanese style) English auction while players can observe whether the other player is still in the process. Players choose at what possible punishment level to drop out. At any level where one of the players decides to drop out first, the process stops for a short time to allow the other player to immediately drop out as well at the same own punishment level. (Of course the other player need not do so and may also choose to drop out later.) Under this mechanism, players can thus reciprocate small pieces of vulnerability step by step. In contrast, in the *Leap mechanism* both players simultaneously submit one final possible punishment level, just as in a sealed bid auction. Here a player thus takes a leap in the dark when choosing to make himself vulnerable.

After the first stage is completed, the resulting possible punishment levels – denoted x_1,$x_2 \geq 0$ – are revealed to both players. Players subsequently play the following prisoner's dilemma game in the second stage:

[6] Sutter et al. (2010) discuss more studies in which players choose between existing insitutions.

[7] Taking quite a different approach Rtischev (2011) theoretically studies vulnerability from an evolutionary perspective and finds that (under certain assortative matching conditions) it can evolve as strategic trait that fosters cooperation.

Table 1: Payoff matrix of the prisoner's dilemma

	C	D
C	c, c	s, t
D	t, s	d, d

where $t > c > d > s > 0$.[8] Moreover, to facilitate the exposition of the equilibrium analysis, we also assume that $t - c \leq d - s$.

Having observed the behavior in the previous two stages players decide in the third stage whether they want to punish their partner or not. Punishment leads to a fixed cost of $p > 0$ for the punisher, which is assumed to be a small amount ($p < d$). Players can only punish their partner with the possible punishment level the partner determined herself in the first stage.

Here, we focus on symmetric prisoner's dilemma games for simplicity. However, our results can readily be extended to asymmetric prisoner's dilemma games, as we will explain below.

3.2 Equilibrium predictions

We simplify the equilibrium analysis by analyzing the model in reduced form. In particular, we assume that players use the following punishment strategy: cooperators are never punished, while defectors are punished with probability one after outcome $C - D$ and with probability $\beta > 0$ after outcome $D - D$. It is also assumed that players punish only if the partner's punishment level is strictly higher than zero, i.e. players are unwilling to pay the punishment cost p if punishment does not cause any harm. Given assumed punishment behavior, players maximize their own payoffs.

Our ad hoc assumptions regarding punishment behavior are well in line with existing experimental evidence. People tend to punish (much) more if they cooperated but were betrayed, than if they both defected. Furthermore, people usually do not punish others who were nice to them and cooperated.[9] Models of other regarding preferences, like e.g. Rabin (1993), Levine (1998), Fehr and Schmidt (1999)[10] and Bolton and Ockenfels (2000), may provide a behavioral justification for our reduced form approach.[11]

[8] In our notation, c gives the payoffs when both players cooperate and d when both defect. Parameter t gives the 'temptation' payoff to defect when the other cooperates, while s reflects the 'sucker' payoff when a player is the only one cooperating.

[9] These types of behavior have already been observed in other public good game experiments; see e.g Fehr and Gächter (2000) or Sefton et al. (2007).

[10] See Appendix A.3 for an analysis of the inequity aversion model.

[11] Alternatively, a more direct approach as in Andreoni and Samuelson (2006) could be taken. Studying a prisoner's dilemma setting, they make direct assumptions on players' preferences for taking a given action (either cooperate or defect) that rationalize the common findings observed in experiments: viz. that people sometimes prefer to cooperate themselves, but differ in the strength of this preference, and value cooperation more when the other player cooperates as well.

Based on the assumed punishment behavior, we can collapse the second and the third stage into the following simultaneous move game:

Table 2: Expected payoff of the whole game given possible punishment levels

	C	D
C	c, c	$s - p \cdot I_{\{x_2>0\}}$, $t - x_2$
D	$t - x_1$, $s - p \cdot I_{\{x_1>0\}}$	$d - \beta(x_1 + p \cdot I_{\{x_2>0\}})$, $d - \beta(x_2 + p \cdot I_{\{x_1>0\}})$

Here $I_{\{x_i>0\}}$ denotes the indicator function, equal to one if $x_i > 0$ and zero otherwise.

We first consider equilibrium behavior in the above collapsed (second stage) game. Let $\delta_i(x_i, x_j) = Pr(\ i \text{ chooses } D \mid (x_i, x_j))$ denote the probability with which player i defects, given possible punishment levels (x_i, x_j) chosen in the first stage. The following lemma characterizes players' equilibrium strategies in the second stage. All proofs are relegated to Appendix A.1.

Lemma 1 *If the other player j has chosen $x_j > 0$ in the first stage, the equilibrium strategy of player i (with $i \neq j$) in the second stage is as follows:*

(i) If $x_i < t - c \equiv \underline{x}$, then player i will be a defector, that is, $\delta_i^*(x_i, x_j) = 1$ for all $x_j > 0$;

(ii) If $\underline{x} \leq x_i \leq \frac{d-s+(1-\beta)p}{\beta} \equiv \overline{x}$, then player i will be a conditional cooperator; that is, player i prefers to cooperate if the other player cooperates and prefers to defect if the other player defects;

(iii) If $x_i > \overline{x}$, then player i will be an unconditional cooperator, that is, $\delta_i^*(x_i, x_j) = 0$ for all $x_j > 0$.

If the other player j has chosen $x_j = 0$ in the first stage, then for player i the same strategies hold with $\overline{x} = \frac{d-s}{\beta}$.

Lemma 1 shows that when a player chooses either a very low or a very high punishment level, he has a dominant strategy in the subsequent second stage game. For punishment levels in between $\underline{x}$ and $\overline{x}$ the player becomes a conditional cooperator, and prefers to coordinate to match the other player's choice.[12]

If there is at least one player with a dominant strategy, a unique equilibrium in the (collapsed) second stage subgame results. Only if both players turn out to be conditional cooperators, multiple equilibria exist: $D-D$, $C-C$, and a mixed one. In that case the two players potentially face a coordination problem. Note, however, that it is unlikely that players will end up $D-D$, given that they could have done so more cheaply by choosing $x_i = 0$ (and thus becoming a defector) in the first stage. They would then arrive at the same outcome but save at least

[12]Lemma 1 assumes that if player i is exactly on the border (i.e. $x_i = \underline{x}$ or $x_i = \overline{x}$), he behaves like a conditional cooperator. Obviously, ties between the three cases could have been broken differently, leading to essentially the same results.

$\beta\underline{x}$ in terms of expected punishments. As it appears, besides D-D also mixing cannot occur on the equilibrium path. The intuition here runs as follows. The mixed equilibrium of the second stage subgame yields a player less than outcome $C-C$. A best response to the other player being a conditional cooperator is, therefore, to become an unconditional cooperator oneself by bidding above $\overline{x}$. In that case coordination on $C-C$ is secured and the mixed equilibrium is avoided.

Lemma 1 presents the cutoff values $\underline{x}$ and $\overline{x}$ for symmetric prisoner's dilemmas. For asymmetric prisoner's dilemmas these cutoff levels will differ between the two players. Yet the underlying mechanism that players move from being a defector to conditional cooperator to unconditional cooperator when they increase their own x remains exactly the same.

The following proposition translates the above intuitions to the equilibria of the entire game.

Proposition 1 *In both versions of the mechanism there are multiple subgame-perfect equilibrium outcomes:*

(i) $x_i = 0$, *and* $\delta_i^*(0,0) = 1$ *for* $i = 1, 2$;

(ii) $\underline{x} \leq x_i^* \leq \overline{x}$, *and* $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*);*

In the Gradual mechanism two additional sets of equilibrium outcomes exist:

(iii) $\underline{x} \leq x_i^* \leq \overline{x}$ *and* $x_j^* > \overline{x}$, *and* $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*);*

(iv) $x_i^* > \overline{x}$, *and* $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*).*

The proposition only specifies the behavior on the equilibrium path; off path behavior is characterized in Appendix A.1. Note that there are infinitely many equilibria for both mechanisms. Yet the Gradual mechanism allows a larger set of equilibria, as it also includes equilibria in which either one or both of the players turn themselves into an unconditional cooperator in the first stage (cf. Proposition 1 (iii) and (iv)). This cannot occur in the Leap mechanism, because the best response to the other player choosing $x_{-i}^* > \overline{x}$ is to choose $x_i = 0$ oneself. However, in the Gradual mechanism it may happen, since players can simultaneously increase their own possible punishment levels in small steps (and cannot go back). They thus can secure that they become an unconditional cooperator only if the other player becomes at least a conditional cooperator at the same time. This ensures that outcome $C-C$ is reached.

With multiple equilibria, the question of interest becomes in which equilibrium players will ultimately end up: either in the non-cooperative equilibrium of part (i), or in one of the cooperative equilibria of parts (ii) through (iv) of Proposition 1. Regarding the latter, in the cooperative equilibria of part (ii) the indeterminacy of the cutoff level for x_i^* to end up in the $C-C$ outcome is caused by the fact that if both players are conditional cooperators, multiple equilibria exist in the second stage subgame (as discussed above). Even though on the equilibrium path the players should always end up in the $C-C$ equilibrium,

off the equilibrium path they may coordinate on one of the other second stage equilibria. Depending on this off path behavior, the equilibrium cutoff for x_i^* is higher or lower.[13]

In the Gradual mechanism players can increase their own punishment level step by step while observing whether the other player follows suit. One would therefore expect that coordination on the cooperative outcome should be easier. By using a forward induction like argument that refines the set of equilibria, the following proposition shows that this is indeed the case.

Proposition 2 *In the Gradual mechanism, only the equilibria described in part (iv) of Proposition 1 survive the iterated elimination of weakly dominated strategies.*

The intuition here is that in the Gradual mechanism players can profitably use a 'wait and see' strategy. Instead of jumping out first, a player can simply wait and see what the other player does. If the other player immediately drops out at zero, it is best to immediately follow suit. Otherwise, the players use the mechanism to achieve the cooperative outcome. In the proof of Proposition 2 it is shown that dropping out first below $\overline{x}$ is (iteratively) weakly dominated. Equilibria in which both players stay in until at least $\overline{x}$ are thus more focal.

For the Leap mechanism we cannot reduce the set of equilibria by eliminating weakly dominated strategies.[14] The reason is that there, unlike in the Gradual case, players cannot condition their own punishment level on the punishment level of their partner. They may therefore remain trapped in skeptical beliefs that the other will choose $x_{-i} = 0$, because there is nothing the other player can do to disprove such skepticism. To illustrate, suppose that although player i would like to choose a possible punishment level that makes him a conditional cooperator if the partner does so as well, he prefers to choose zero if his partner chooses zero. Skeptical beliefs then induce player i to choose a punishment level of zero, followed by defection in the second stage. If the other player thinks the same, players are stuck in the non-cooperative equilibrium even though both prefer a cooperative one. In contrast, in the Gradual mechanism players can freely disprove their partner's skeptical beliefs, simply by waiting and not dropping out first at zero.

To sum up, in the Leap mechanism players face strategic uncertainty, as there appears to be no compelling argument that makes coordination on one of

[13] The indeterminacy in parts (iii) and (iv) of Proposition 1 results from the fact that if both players always cooperate, punishments are never carried out. In that case players are indifferent between all possible punishment levels (within the given ranges). As a result, in e.g. part (iv) any x_i^* above $\overline{x}$ can be supported in equilibrium.

[14] Ben-Porath and Dekel's (1992) result is very similar to ours, but obtained for other settings. They show that if players have the possibility to simultaneously burn money before playing a coordination game, the "worse" equlibrium cannot be eliminated by iterated elimination of weakly dominated strategies. Unlike their setup, in our model players do not literally burn money to show their intention for cooperation. Moreover, the strategic nature of our second stage game depends on the possible punishment levels chosen in the first stage: for low possible punishment levels it corresponds to a prisoner's dilemma, while for e.g. in between possible punishment levels it becomes like a coordination game.

the cooperative equilibria focal. Rational players will coordinate on cooperation in the Gradual case, because outcomes that correspond to other equilibria make use of strategies that do not survive the iterated elimination of weakly dominated strategies. Players can safely choose higher possible punishment levels in the Gradual meachanism than in the Leap mechanism. The Gradual mechanism is therefore predicted to perform better in fostering cooperation.[15]

To conclude this section we briefly illustrate how models of other regarding preferences can rationalize the assumed punishment behavior. Take for instance the inequity aversion model developed by Fehr and Schmidt (1999). This model predicts that players who sufficiently dislike being behind, punish the defector after outcome $C-D$ because this reduces inequality.[16] At the same time, the defector will typically prefer not to punish, as this decreases his own payoff. If both players choose the same action in the prisoner's dilemma, multiple punishment equilibria may exist. The intuition is that, depending on the choices made in the first two stages, players may want to mimic the other's punishment behavior. They may then either coordinate on an equilibrium in which both punish, or another one in which they do not. After joint cooperation (outcome $C-C$) no punishment serves arguably as focal point; there is simply no compelling reason why players should punish. Yet after outcome $D-D$ it is much less clear what the focal punishment equilibrium is; one could either argue that punishment is appropriate given that the other defected, or make the case that punishment is superfluous because both players did so. Our β parameter reflects this indeterminacy and can capture the average behavior of inequity averse players in practice. With Fehr-Schmidt style preferences, Lemma 1 and Proposition 1 remain qualitatively valid, except that the thresholds $\underline{x}$ and $\overline{x}$ will then depend also on the inequity aversion parameters of the players involved, and on the other player's chosen possible punishment level. Types (i.e. being a defector or (un)conditional cooperator) are not independent anymore, and they can also change if the partner chooses a different possible punishment level.

4 Experimental design and procedures

The experiment was conducted in the CREED-laboratory at the University of Amsterdam. In total, 144 subjects participated in 6 sessions. None of the subjects participated in more than one session. Subjects were mainly undergraduate students from various fields (e.g. economics, business, psychology, law). At the start of the computerized experiment, subjects received the instructions on their screen. Subjects read the instructions at their own pace and had to successfully answer some control questions testing their understanding before they could

[15] The same general prediction holds when we assume instead that the incentive to defect is stronger if the other player cooperates than when the other player defects ($t-c>d-s$). In Appendix A.2 we provide the full equilibrium characterization for that case and show that coordination on a cooparative outcome then can be sustained as equilibrium outcome in the Leap mechanism only if β is sufficiently low. In contrast, equilibrium cooperation can still occur under the Gradual mechanism for all values of β.

[16] For the formal analysis see Appendix A.3.

proceed with the experiment. The instructions for one of the treatments are included in Appendix B.[17] Because the games in two of the three treatments are not simple, we used meaningful labels like punishment levels, cooperation and defection in the instructions of all treatments.

Subjects received a starting capital of 500 points at the beginning of the experiment, and they could earn additional points by their decisions. At the end of the experiment, their points were converted into real money; the conversion rate was such that 100 points corresponded to 1 euro. In an experiment that lasted between 1 to 2 hours, subjects earned on average 19.7 euros (with a minimum of 12.9 euro, and maximum of 29.6 euro).

There were 24 subjects in each session. In each of the 50 rounds, subjects were randomly rematched in pairs within their matching group of 8 subjects. Subjects were informed that they would never meet twice in a row with the same other subject. We organized two sessions for each of the three treatments. Thus, for each treatment we collected data of 6 independent matching groups.

We varied the trust-building mechanism between treatments. In the Control treatment, there was no mechanism, and the game consisted of only one stage in which the two subjects that formed a pair in a round played a prisoner's dilemma. Both subjects simultaneously chose between Cooperate (C) and Defect (D). At the end of the period, subjects were informed of each other's choices and received a payoff corresponding to the action pair. The payoffs in the prisoner's dilemma are listed in Table 3.

Table 3: Payoffs in the prisoner's dilemma in the experiment

	C	D
C	$55; 55$	$5; 70$
D	$70; 5$	$25; 25$

The other two treatments allowed for trust-building as in the model. In each of these treatments, each round had a three-stage structure. In the first stage, subjects determined their own possible punishment levels. After the first stage, subjects were informed about their partner's possible punishment level. In the second stage, subjects played the prisoner's dilemma game listed in Table 3. At the end of this stage, they were informed about their partner's action in the prisoner's dilemma. In the third stage, subjects decided whether they wanted to punish their partner or not. The possibility to punish the partner was not limited to cases where the partner defected. If they wanted, they could also punish a partner who cooperated. If a subject decided to punish, the partner received a deduction in points equal to his own possible punishment level of the first stage. At the same time, the subject who decided to punish incurred a cost of 4 points. After the third stage, punishment decisions of the own pair and earnings were revealed to the subjects. In each round, a subject's earnings was equal to the payoff in the prisoner's dilemma game, possibly diminished by the

[17] The instructions for the other two treatments are similar and will be sent upon request.

received punishment and/or the punishment cost (if punishment was received and/or given).

The two trust-building treatments differed in the way in which trust could be built in the first stage. In treatment "Gradual", an automatized clock slowly raised the possible punishment levels, and subjects could stop the clock at the desired integer own possible punishment level. If one subject in a pair submitted a punishment level, the partner was immediately informed about this. In order to give the second subject the chance to submit the same punishment level as the first one, the clock was stopped for a few moments after the first subject had stopped. The clock only stopped for a given pair, not for everybody in the experiment. If the second subject decided not to stop at the same level, the automatized clock continued to rise until the second subject was satisfied and stopped the clock. To guarantee that 50 rounds could be run in a reasonable time frame, we imposed a maximum possible punishment level of 50 in this stage. When the clock reached this amount, subjects who still remained in the process automatically dropped out of the mechanism, and their possible punishment level was set equal to this maximum amount. Subjects were aware of this procedure.

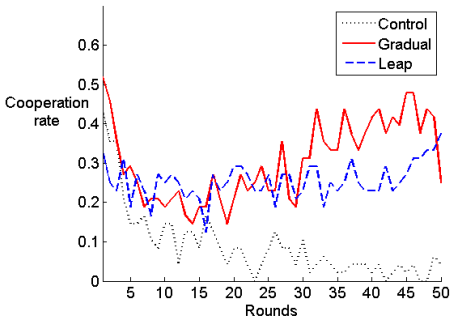
In the first stage of treatment "Leap", subjects essentially had to make a leap in the dark and choose an own possible punishment level without having any information about the partner's own possible punishment level. Subjects simultaneously chose the own possible punishment level, a single integer number of at least zero and at most the maximum possible punishment level of 50. In the other stages, there were no further differences between Gradual and Leap.

With the parameters used in the experiment, players can turn themselves into a conditional cooperator if they choose an own possible punishment level of at least 15. The upper bound of the conditional cooperator range depends on β, and is strictly decreasing in it. This upper bound is at least $d - s = 20$, and it goes to $+\infty$ as β goes to 0. Above this threshold, players turn themselves into unconditional cooperators.

In all three treatments, a social history screen with the 10 most recent completed rounds was always visible. In Gradual and Leap, this screen contained the possible punishment levels, the actions, and the punishment decisions of each pair in the subject's matching group. The observations were ordered on the basis of the own possible punishment level of the first player in a pair.[18] Appendix B provides an example of what a social history screen may look like. We decided to give this social history screen because the game is complex and we wanted to facilitate better understanding and speed up learning. Furthermore, the history screen helps coordinating on a particular equilibrium in the game. In the social history screen of the control treatment we provided information about the actions of the pairs in a matching group.

[18] If player 1's own punishment levels were the same across pairs, the ordering was determined by the own possible punishment level of the player 2; if these numbers were also the same, then they were sorted on the basis of the cooperation decisions. It was randomly determined who in a pair was listed as player 1 and who was listed as player 2.

Figure 1: Cooperation rate over time



5 Results

In Section 5.1 we present an overview of the experimental results. We start with the question how the achieved outcomes in the prisoner's dilemma vary with the treatments. Then we analyze how individuals behave in the three stages of the treatments that use one of the mechanisms to build trust. First we deal with the extent to which subjects make themselves vulnerable in stage 1. Then we show how the possible punishment levels of stage 1 map into decisions in the prisoner's dilemma in stage 2. Finally we discuss how actual punishment behavior in stage 3 depends on the behavior of stage 2. In Section 5.2 we zoom in on the dynamics in our data, and we provide an explanation of the results.

5.1 Overview of the experimental results

We start with the question whether the mechanism works as predicted. That is, do subjects cooperate more often when a mechanism is used and if so, is Gradual more successful than Leap in achieving this goal. Figure 1 displays the cooperation rates over time in the three treatments. In agreement with our conjecture that we employed an environment in which it is hard to sustain cooperation, cooperation levels fall dramatically in the Control treatment. In the second half of the experiment, cooperation rate is around 4%. Interestingly, in the Gradual treatment cooperation levels also fall rapidly in the beginning of the experiment. Apparently it takes time before the mechanisms become effective. After round 25 the cooperation rate grows to approximately 26.5% in Leap and 36.25% in Gradual, while it falls short of 4% in Control.

Table 4 presents the cooperation rates, and also the levels of the proportions of $C-C$ and $C-D$ outcomes and shows to what extent the differences between treatments are statistically meaningful. For the non-parametric test results reported in this paper, we employ a prudent testing procedure in which we use average statistics per matching group as data points. Panel A of the table focuses on a comparison of the cooperation rates between treatments. The main message here is that the differences between each of the treatments that

use a trust-building mechanism and the control treatment are significant in the second half of the experiment, while the difference between Gradual and Leap is never significant. The latter result is caused by the fact that not all matching groups in Gradual actively exploit the possibilities offered by the mechanism. The difference between Gradual and Leap lies more in the extent to which the mechanism is used if it is used. We come back to this in Section 5.2.

Although the increased cooperation rates already show that the mechanism is successful to foster cooperation, it is also important to examine how the mechanism leads to an increase in ending up in the cooperative outcome. To answer this question, panel B and C of Table 4 presents the proportion of the $C-C$ and $C-D$ outcomes. From panel B it can be seen that subjects ended up in the $C-C$ outcome significantly more often in both treatments with the mechanism compared to the Control treatment, both in the first and in the second half of the experiment, while again, the difference between the Gradual and Leap treatment is never significant. Panel C of the table presents the comparison of the proportions of $C-D$ outcomes. The proportion of $C-D$ outcomes diminishes over time in the three treatments. In the first half of the experiments, $C-D$ outcomes are observed approximately equally often in the three treatments, and all pairwise differences between the three treatments are not significant. In the second half of the experiment, the rate of $C-D$ outcomes drops substantially in Gradual and Control, but much less so in Leap. Remember that Gradual offers better opportunities than Leap to prevent $C-D$ outcomes, because in the former subjects may use the first stage to coordinate their possible punishment levels. In agreement with this intuition, subjects appear to be more successful in preventing $C-D$ outcomes in Gradual than in Leap. As a result, (weakly) significantly less $C-D$ outcomes are observed in Gradual than in Leap in the second half of the experiment.

We now turn to an analysis of individual behavior in the three stages of Gradual and Leap. We first deal with the extent to which subjects are willing to make themselves vulnerable in stage 1. Figure 2 illustrates how average own possible punishment levels change over time for each mechanism. Average possible punishment levels increase as time passes for both treatments. However, this increase is larger in Gradual than in Leap. Figure 1 and Figure 2 together suggest that subjects first try to achieve mutual cooperation without making use of the mechanism. Then when they find out in approximately round 10 that this does not work very well, they start using the mechanism which gradually enhances cooperation levels.

In the remainder of this section, we focus on the second half of the data. The reason is that we want to compare the treatments in the phase of the experiment in which subjects have learned about the potential usefulness of the mechanism in the game that they are playing. In Section 5.2 we come back to the data of the whole experiment when we describe the dynamics in the data. Figure 3 shows the distribution of possible punishment level pairs for each mechanism. The distributions of punishment level pairs differ substantially across treatments. As expected, possible punishment levels are much closer to each other in Gradual than in Leap; note that pairs of possible punishment

Table 4: Percentage of cooperation, $C-C$ and $C-D$ outcomes

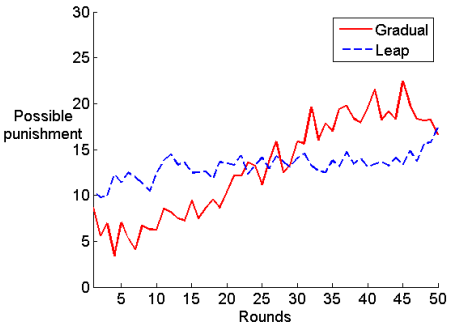
Treatment	First half (rounds 1-25)	Second half (rounds 26-50)
Panel A. Cooperation rate		
Control	13.67%	4.25%
Gradual	24.67%	36.25%
Leap	24.50%	26.50%
Control vs. Gradual	0.15	0.09*
Control vs. Leap	0.17	0.04**
Gradual vs. Leap	0.69	0.63
Panel B. C-C outcome		
Control	3.16%	0.50%
Gradual	13.50%	33.67%
Leap	13.67%	18.67%
Control vs. Gradual	0.04**	0.01**
Control vs. Leap	0.09*	0.05**
Gradual vs. Leap	0.63	0.26
Panel C. C-D outcome		
Control	21.00%	7.50%
Gradual	22.33%	5.17%
Leap	21.67%	15.67%
Control vs. Gradual	0.81	0.57
Control vs. Leap	0.94	0.11
Gradual vs. Leap	0.87	0.06*

Notes: **: significant at 5% level, *: significant at 10% level according to ranksum test with $n = 6$. The proportions of the given outcome and the p-values for tests for differences across treatments are displayed in the first and the second half of each panel, respectively.

levels are mainly along the diagonal in Gradual. There, subjects rarely choose a higher possible punishment level after they have been informed of their partner's possible punishment level. In Leap, subjects do not have the possibility to condition their possible punishment level on the level chosen by the partner, thus these levels correlate less. Overall, the correlation between the players' possible punishment levels equal 0.89 in Gradual and 0.35 in Leap.

The figure shows that in both Gradual and Leap a substantial proportion of pairs of subjects coordinate on low levels of possible punishment (0, or in the interval [1,5]). Another interesting feature of the figure is that both distributions have, besides the origin, another but different spike. In Gradual, subjects often choose a very high possible punishment level (in the interval [46;50]), while in Leap, many pairs manage to coordinate on a possible punishment level in [16;20]. The locations of these alternative spikes are in agreement with the theory presented in Section 2. If we condition on cases where the minimum possible punishment level of a pair is at least 15 (the theoretically threshold above which cooperation becomes possible), the average punishment level in Gradual (39.5) is almost twice as large as in Leap (20.6). The difference is significant according to a Mann-Whitney test (p=0.03). Thus, when pairs of subjects aim for cooperation, they choose substantially higher possible punishment levels in Gradual than in Leap.

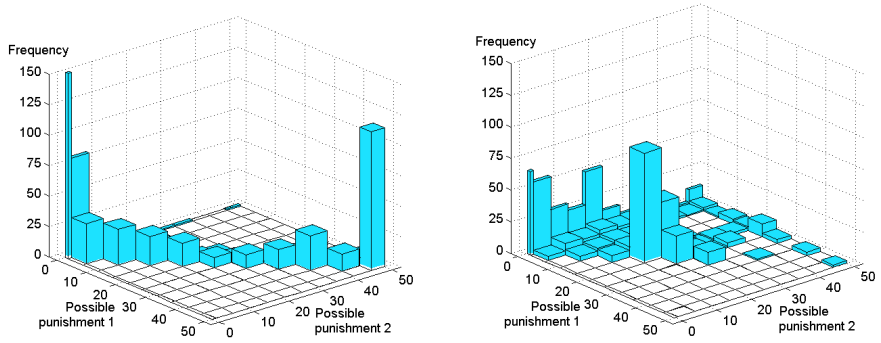
Figure 2: Evolution of average punishment levels for each mechanism



Now we consider how the chosen possible punishment levels affect actions in the prisoner's dilemma. Figure 4 shows how often subjects cooperate for given combinations of possible punishment levels. The darkness of a circle corresponds to the extent to which subjects cooperate for a combination of the own and the partner's possible punishment level; the darker the circle is, the more likely it is that subjects choose to cooperate. The size of the circle indicates how often a pair of possible punishments is observed in a treatment. The following pattern emerges from the figures. Only when their own possible punishment level surpasses 15, subjects seriously consider to cooperate. In the quadrant where both players' own punishment levels exceed 15, subjects gradually cooperate more often when their own punishment level increases as well as when the other's own punishment level increases. The former result is in line with the notion that very high own possible punishment levels turn a player into an unconditional cooperator. The latter result agrees with the possibility that if a player chooses an own punishment level in the conditional cooperator range while the partner has become an unconditional cooperator, the player can become more confident that cooperation is the better choice. Interestingly, even in the quadrant where a player's own punishment level exceeds 15 while the partner's level falls short of 15, players tend to cooperate when they choose very high own punishment levels. This result supports the idea that with unilateral very high own punishment levels players face the danger of becoming an unconditional cooperator who can be exploited by the partner. This danger only materializes occasionally in Leap.

In a logit model, we investigate more carefully how possible punishment levels map into decisions in the prisoner's dilemma. Table 5 reports the results of the logit regressions with fixed effects, in which we employ the player's id as panel variable. We cluster the standard errors by matching groups, because subjects' decisions are not independent from those with whom they are matched. Since we have a between subjects design and estimate a fixed effect model where the coefficients are only identified through the "within" dimension of the data (that is within subjects), we cannot directly estimate whether being in Gradual

Figure 3: Distribution of punishment levels in the Gradual (left panel) and in the Leap treatments (right panel)



Notes: We define 11 categories for the 50 possible punishment levels. The first category equals the singleton set {0}, because many subjects choose that level. The other 10 categories cover the range from 1 to 50 with equal category length (that is, [1;5], [6;10], and so on). Each possible punishment level in each pair is assigned to a category, and pairs are plotted according to these categories. The figure is based on rounds 26-50. We plot the punishment levels of all pairs such that punishment level 1 $\leq$ punishment level 2 (thus, by definition, there are no observations in the front part of the histogram).

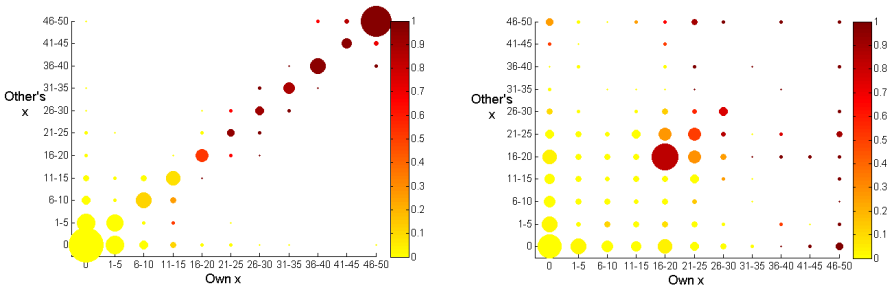
has a different effect on the decisions than being in Leap. Instead, to identify a possible treatment effect we include interaction terms between treatment and the main variables of interest. In all three models we regress the cooperation decision (cooperate: 1, defect: 0) on (i) a dummy variable which is 1 if both players choose a possible punishment level at least 15, otherwise 0, (ii) the own possible punishment level if it is at least 15 and the other player's is lower than 15, and (iii) the interaction between treatment and the first dummy variable.[19]

The first model does not include more variables. Its estimation result is that cooperation is significantly more likely if both players choose a punishment level of at least 15, that is, they both turn themselves into a conditional cooperator, as theory predicts. Further, when a subject's own punishment level exceeds the threshold of 15 while the partner's falls short of it, the subject is significantly more inclined to contribute when the own possible punishment level increases, in agreement with the idea that high own punishment levels may make players committed to cooperate. The second model additionally includes the other player's possible punishment level if both players choose a punishment level at least 15, and its interaction term with the treatment. When they have become conditional cooperators, subjects contribute significantly more often when the other chooses a higher own punishment level. Finally, in the third model we also add the own possible punishment level if both players choose a punishment level at least 15.[20] The own punishment level also has a significant and positive

[19] We do not include an interaction term between treatment and the second variable because there are only a few observations in that region in the Gradual treatment.

[20] Here we do not add the interaction term with treatment, because this interaction term is

Figure 4: Cooperation probability for a given punishment level pair in the Gradual (left panel) and in the Leap treatments (right panel)



Notes: The 11 categories for the 50 possible punishment levels are the same as for Figure 3.

effect on the cooperation probability if both punishment levels are at least 15. This again suggests that subjects may become more inclined to cooperate when they have turned themselves into unconditional cooperators.[21]

It is important to note that the interaction terms with the treatment dummies are never significant which shows that in Gradual and Leap subjects respond rather similarly to given combinations of possible punishment levels.

Finally we discuss the question which choices in the prisoner's dilemma actually trigger punishments. The data are roughly in line with the assumptions of the theoretical model. For a start, own punishment levels of 0 rarely trigger punishments ($< 1\%$ of the cases). Conditional on strictly positive own punishment levels, Table 6 presents for each treatment how often subjects punish after a combination of actions in the prisoner's dilemma. Subjects primarily choose to punish after they cooperated while the partner did not (that is, in the $C - D$ outcome). These punishment frequencies are close to our theoretical assumption that the cooperator punishes with probability 1 if the other defects. In either treatment, cooperating subjects punish significantly more often when the partner defects than when the partner cooperates. If we combine all cases where the partner cooperates, subjects punish in only 8% of the cases in Gradual and in 5% of the cases in Leap. These numbers are not substantially higher than zero, providing support to our assumption that a cooperative partner is not punished. When both subjects defect, punishments occur, but only in a minority of the cases. Pooling the data of the two treatments, this occurs in

highly correlated with the interaction term with the other's possible punishment level if both players choose a punishment level at least 15.

[21] In the reported regressions, we use the theoretical threshold of 15 as the lowerbound of the range in which players can turn themselves into conditional cooperators. We also ran the 3 regressions for any threshold level between 10 and 20. It turns out that the likelihood for model I is maximized for a threshold at 17 and for models II and III at threshold 16. Notice that with an own punishment level of 15 subjects are indifferent between cooperating and defecting if the partner cooperates. It turns out that subjects cannot turn themselves credibly into conditional cooperators with lower own punishment levels, which might have been the result if subjects dislike being ahead of the other.

Table 5: Logit model for the cooperation probability

Dependent variable: Cooperation	Model I	Model II	Model III
$I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$	5.84 (1.31)***	2.11 (2.29)	-0.78 (2.77)
$x_{\text{own}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} < 15\}}$	0.09 (0.03)***	0.12 (0.06)**	0.12 (0.05)**
$x_{\text{own}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$	-	-	0.15 (0.06)**
$x_{\text{other}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$	-	0.22 (0.06)***	0.20 (0.05)***
Gr*$I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$	-0.42 (1.77)	-3.92 (2.95)	-1.15 (3.34)
Gr*$x_{\text{other}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$	-	0.07 (0.12)	-0.05 (0.14)
-Log likelihood	372.77	268.26	255.86
Number of panels (subjects)	69	69	69
Number of observations per panel	25	25	25

Notes: ***: significant at the 1% level, **: significant at 5% level, *: significant at 10% level. In Model I we regress cooperation decisions (cooperate: 1, defection: 0) on a dummy variable which is 1 if both players choose a possible punishment level higher than 15 ($I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$), on the own possible punishment level provided that it is higher than 15, and the other's is lower than 15 ($x_{\text{own}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} < 15\}}$), and on an interaction term with the first variable and a treatment dummy (Gradual: 1, Leap: 0). Model II includes the previous regressors, and the other's punishment level provided that both players chose a punishment level higher than 15 ($x_{\text{other}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$) and an interaction term between this variable and the treatment dummy. Model III includes Model II, and the own punishment level provided that both players chose a punishment level higher than 15 ($x_{\text{own}} * I_{\{x_{\text{own}} \geq 15\}} * I_{\{x_{\text{other}} \geq 15\}}$) as an additional regressor. Regression based on second half of the experiment (rounds 26-50). Std errors are in parentheses after the coefficients (Std errors are adjusted for 11 clusters).

approximately 13% of the cases. The frequency of punishment after $D - D$ increases to approximately 20% if we condition on cases where a subjects's own punishment level is larger than 15.

Roughly the same punishment pattern emerges in the two treatments where punishments are allowed. In both cases, subjects most often punish if they cooperate and the partner defects, while they least often punish in cases where they and their partner cooperate. For none of the four possible action combinations in the prisoner's dilemma, the actual punishment probability differs significantly across treatments.

So far the following picture emerges from our data in Gradual and Leap. In both treatments, subjects either refrain from making themselves vulnerable or they make themselves vulnerable to a sizable extent. If they refrain from making themselves vulnerable, subjects primarily choose to defect. On the other hand, if they do make themselves vulnerable, they tend to do this to a substantially larger extent in Gradual than in Leap. The process of mutually monitoring allows subjects to coordinate on high own punishment levels in Gradual. Given a combination of own possible punishment levels, subjects choose similar actions in the prisoner's dilemma in both treatments. The higher the own possible punishment level is (given that it is already at least 15), the larger the probability of cooperation is. In addition, subjects' cooperation decisions respond positively to the other's possible punishment level. Finally, we do not observe systematic differences in the treatments in how combinations of choices in the prisoner's dilemma map into actual punishments. Therefore, the main difference between the treatments is that subject make themselves vulnerable to a larger extent in

Table 6: Punishment behavior

Own action	Treatment	Partner's action		p
		Cooperate	Defect	
Cooperate	Gradual	0.06 (404)	0.93 (29)	0.04
	Leap	0.02 (220)	0.88 (78)	0.07
	p-value	0.62	0.13	
Defect	Gradual	0.37 (30)	0.14 (316)	0.89
	Leap	0.13 (88)	0.13 (472)	0.35
	p-value	0.45	0.52	

Notes: Percentage of given punishment (excluding cases with partner's punishment level zero) with the number of cases in parentheses. The p-value shows the test results of testing for the differences across treatments based on the Mann-Whitney test, whereas Wilcoxon p compares behavior within a treatment.

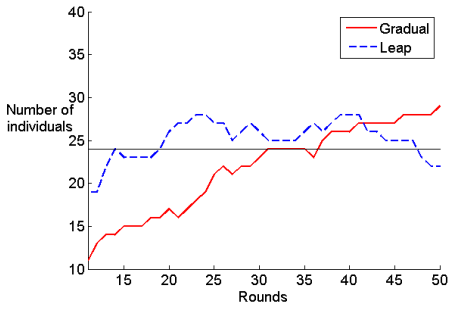
Gradual if they aim for cooperation. This feature of our data is in line with the theoretical analysis. Theoretically, mutual high own possible punishment levels can be supported in Gradual by a process of iteratively eliminating weakly dominated strategies. Such a process is impossible in Leap, where the best that subjects can do is to turn themselves into conditional cooperators. From this perspective, it makes a lot of sense that subjects choose higher own possible punishment levels in Gradual than in Leap.

5.2 Dynamics in the data and explanation of main result

In this section, we take a closer look at the dynamics in our data. In the previous section, it was shown that the positive effect of the trust-building mechanisms primarily emerges in the second half of the experiment. Even then, it is not the case that all subjects switch to using the mechanisms. To get a sharper view on how and why play evolves as it does in our experiment, we classify each subject as user or non-user. We do this on the basis of a subject's behavior in the preceding 10 rounds.[22] In Leap, we say that a subject is a user if the own possible punishment levels are 15 or higher in more than 5 times of the preceding 10 rounds. We employ 15 as a threshold to judge an own possible punishment level in a round because theoretically from this level players credibly signal that they are interested in achieving the cooperative outcome. In Gradual, the characterization is not as simple, because subjects who want to use the mechanism may revert to a low own possible punishment level after they find out that the partner chooses a low level. However, also in Gradual there are certain types of behavior that unambiguously characterize a user or a non-user. If a subject chooses an own possible punishment level of 15 or higher, the subject behaves as a user. If a subject chooses the own possible punishment level earlier than the partner, while the possible punishment level has not yet reached 15, the subject behaves as a non-user. Only when a subject submits a low punishment

[22]It follows from this choice that we cannot characterize subjects in the first 10 rounds, but only from the 11^{th} one onwards.

Figure 5: Number of users of the mechanism



level (lower than 15) after the partner decided on the own punishment level, we cannot use the case to classify the subject because we do not know if the subject had been willing to stay in until a level of 15 or higher if the partner had done the same. When classifying a subject on the basis of the preceding 10 rounds, we simply ignore the latter cases and say that a subject is a user if and only if the subject behaves as a user in more than 50% of the remaining cases.[23]. Note that our procedure allows for the possibility that a subject's type changes over time.

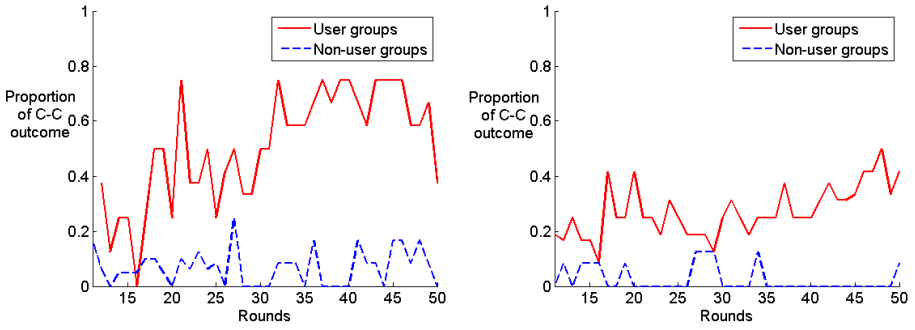
Figure 5 illustrates how the relative frequency of types evolves over time in the two treatments where the mechanism can be used. The horizontal line at 24 corresponds to 50% of the subjects (in each treatment we have observations of 48 subjects). In Gradual the number of users increases considerably over time (starting from 11 users), and after round 37 there are more users than non-users. In contrast, in Leap the number of users starts at a higher level and increases less steeply. In either treatment the frequency of users seems to converge to a small majority of the subjects.

A trust-building mechanism is only useful if other subjects in a matching group employ it as well. In each round, we classify a matching group of 8 subjects as a user group if it has at least 4 users in the given round. In Leap, there are 4 user groups in more than 60% of the rounds (in the remaining rounds there are 3 user and 3 non-user matching groups). In contrast, there are 4 or 5 non-user matching groups till round 25 in the Gradual treatment, and there are almost always 3 user and 3 non-user groups in the second half of the experiment. This shows that the better performance of Gradual does not result from an increase in the proportion of subjects or the proportion of groups that actually adopt the mechanism.

Instead, it appears that user groups in Gradual are more successful than user groups in Leap. Figure 6 shows the evolution of the percentage of cooperative $C - C$ outcomes over time for each treatment, separately for user and non-user groups. Unsurprisingly, user matching groups end up more often in the

[23]If a subject's behavior does not identify the type in any of the 10 preceding rounds, we keep the type assigned in the previous round.

Figure 6: Proportion of $C-C$ outcomes in the matching groups in the Gradual (left panel) and in the Leap treatments (right panel)



$C-C$ outcome than non-user groups in either treatment. Furthermore, there is no big difference between the efficiency of non-user groups across treatments. Remarkable is the substantial difference between user groups across treatments. In the second half of the experiment, the frequency of $C-C$ outcomes per user group is approximately twice as high in Gradual.

A plausible explanation is suggested by the observation that users choose higher possible punishment levels in Gradual than in Leap.[24] As we saw from the regressions of cooperation decisions (cf. Table 5), higher possible punishment levels above 15 significantly increase the likelihood of cooperation. These two observations together explain our finding that the Gradual mechanism is more efficient in facilitating coordination on the cooperative outcome.

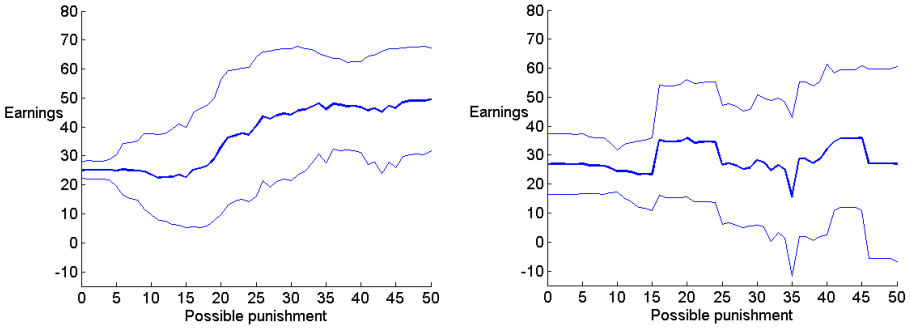
Finally we shed some more light on the question of why subjects in Gradual choose higher own possible punishment levels and on the question of why not everybody in a treatment uses the mechanism. For the second half of the experiment, Figure 7 shows how subjects' earnings depend on their own possible punishment level. The figure displays per treatment the average earnings as function of the own possible punishment levels, as well as the standard deviations around these averages.

In Gradual average earnings remain approximately constant until the possible punishment level of 15 which represents the theoretical threshold for cooperation to occur. From this level, average earnings steadily increase with the own punishment level. The picture is quite different for Leap; there subjects' earnings are already maximized for intermediate own punishment levels in the range of 15-25. So given the strategic uncertainty that subjects face, they are well advised to choose lower own possible punishment levels in Leap than in Gradual, just like they do.

In both treatments, the standard deviations around the averages increase

[24] It is not the case that groups are more homogeneous in Gradual. We do not observe systematic differences in the treatments in how often a homogeneous group of at least 6 users formed.

Figure 7: Average earnings for the own possible punishment levels in the Gradual (left panel) and in the Leap treatments (right panel)



Notes: The data is smoothed by using weighted moving averages where the weights are the number of observations for the given own possible punishment level.

with the possible punishment levels. So subjects face a clear trade-off; if they want to maximize expected payoff, they have to use the mechanism and choose very high own possible punishment levels in Gradual or intermediate own punishment levels in Leap. With this behavior they face higher risk than when they completely refrain from using the trust-building mechanism and set the own possible punishment levels equal to 0. This observation sheds light on the fact that not all subjects in a treatment switch to using the mechanism. Possibly some risk averse subjects settle for lower expected earnings with lower risk by refraining from using the mechanism.

6 Conclusion

A common finding in the psychological literature is that strangers often build relationships by the reciprocal disclosure of secrets. This process has consistently been shown to enhance feelings of intimacy. In this paper, we investigate the possibility that cooperation is fostered through a process of enhancing one's own vulnerability. By doing so, a player signals to his partner that he aims for cooperation, and provided that the partner is willing to punish an uncooperative act, the signal becomes credible. We consider the case that people can build trust conditionally in small incremental steps as well as the case in which this has to be done in a single unconditional leap of faith. Theoretically, cooperation can be supported in both cases. At the same time, a simple argument based on the iterated elimination of weakly dominated strategies suggests that higher levels of own vulnerability will be observed in the case where players can gradually build trust.

In the experiment, we observe that after some initial aversion players start using the possibility of building trust through the enhancement of own vulnerability. In the second half of the experiment, they do this to a similar extent

in either variant of the trust-building process. Subjects coordinate substantially more often on the cooperative outcome with a trust-building mechanism than in the control where this is not allowed. So the trust-building mechanism works, albeit not perfectly. In this sense, its performance is comparable to other mechanisms such as Varian's (1994) compensation scheme.

In agreement with theory, we observe higher vulnerability levels when players can build trust gradually than when they have to do it in one leap. In the experiment, subjects' willingness to cooperate increases steadily with the own punishment levels after the threshold level needed for cooperation is reached. Thus, when subjects make use of the mechanism, higher mutual cooperation levels are achieved when trust is built gradually. At the same time, the gradual variant helps reducing the number of miscoordination outcomes where one partner cooperates and the other defects. These findings shed new light on why a slow, gradual process of building trust may be preferable.

Appendix A

A.1 Proofs

Proof of lemma 1:

(i) If $x_i < \underline{x}$ and $x_j > 0$, we have that both $c < t - x_i$ and $s - p < d - \beta(x_i + p)$, from which it follows that player i is better off by defecting, no matter what player j does.

(ii) If $\underline{x} \leq x_i \leq \overline{x}$ and $x_j > 0$, it holds that $c \geq t - x_i$ and $s - p \leq d - \beta(x_i + p)$. Player i then prefers to choose the same action as the other player does.

Case *(iii)* and the proof for $x_j = 0$ go analogously. □

Proof of proposition 1:

We consider the two versions of the mechanism separately.

(a) Leap mechanism. We first show that there is no equilibrium in which $0 < x_i < \underline{x}$ or $x_i > \overline{x}$. If $0 < x_i < \underline{x}$, choosing D is a dominant strategy for player i. Realizing this, the other player j will choose D when $x_j \leq \overline{x}$ and C when $x_j > \overline{x}$. Player i is then punished for his defection with at least probability $\beta > 0$. He is therefore better off if he chooses $x_i = 0$ (and subsequently defect); this would not alter the other player's choice between C and D (which is still governed by $x_j \gtreqless \overline{x}$), but avoids costly punishment. In case $x_i > \overline{x}$ player i will always cooperate, for all values of $x_j \geq 0$ (cf. Lemma 1). The best reply of the other player j is then to choose $x_j = 0$ and defect. But against this strategy of player j it is not a best response for player i to choose $x_i > \overline{x}$ and C; he can earn more by choosing $x_i = 0$ and defect as well.

Two possibilities therefore remain: $x_i = 0$ and $\underline{x} \leq x_i \leq \overline{x}$. Suppose first that $x_i = 0$. Player i then defects for sure in the second stage.(cf. Lemma 1). Player j therefore obtains s by choosing C and $d - \beta x_j$ by choosing D. From $d > s$ she can get the most if she chooses $x_j = 0$ and D as well. Therefore, $x_1 = x_2 = 0$ followed by $D - D$ constitutes an equilibrium. (Off the equilibrium

path behavior is governed by Lemma 1. A unilateral deviation of player i does not affect $j's$ behavior. With j being a defector, $i's$ response is fully determined by the lemma.)

Next suppose that $\underline{x} \leq x_i \leq \overline{x}$. From the above this necessarily requires that $\underline{x} \leq x_j \leq \overline{x}$ as well (note that against $x_j = 0$, choosing $\underline{x} \leq x_i \leq \overline{x}$ is not a best response). Both players are thus conditional cooperators. The (collapsed) second stage subgame then allows three equilibria: $C-C$, $D-D$, and a mixed equilibrium. The cooperative outcome $C-C$ strictly Pareto dominates the other two. Suppose that players coordinate on the mixed equilibrium. Then by deviating to $x_i > \overline{x}$ and making herself an unconditional cooperator (while player j remains a conditional cooperator), player i can secure outcome $C-C$. Therefore, the mixed equilibrium of the second stage subgame cannot occur on the equilibrium path of the entire game. For the same reason this also applies to outcome $D-D$. (Note that, besides $x_i > \overline{x}$, then also $x_i = 0$ would be a profitable deviation.) We thus obtain an equilibrium in which $\underline{x} \leq x_1^*, x_2^* \leq \overline{x}$ followed by $C-C$. Off the path behavior is such that after any $(x_1, x_2) \neq (x_1^*, x_2^*)$ with $\underline{x} \leq x_1, x_2 \leq \overline{x}$, both players choose D. This supports that players choose exactly (x_1^*, x_2^*) in the conditional cooperator range. Off path behavior outside the conditional cooperator range is again governed by Lemma 1.

(b) Gradual mechanism. Also in this case there is no equilibrium in which $0 < x_i < \underline{x}$. To see this, from Lemma 1 it follows that player i defects after choosing such punishment level. The other player j can then only achieve $\max\{s-p, d-\beta(x_j+p)\}$. For $x_j < \overline{x}$ it holds that $\max\{s-p, d-\beta(x_j+p)\} = d-\beta(x_j+p)$, which is decreasing in x_j. Player j's best response is thus to drop out immediately after player i (such that $x_j = x_i$) and defect as well. Choosing $0 < x_i < \underline{x}$ thus yields player i a payoff equal to $d-\beta(x_i+p) < d$, strictly less than choosing zero would do.

Two possibilities therefore remain: $x_i = 0$ and $x_i \geq \underline{x}$. If $x_i = 0$ the same reasoning as for the Leap mechanism applies; choosing $x_i = 0$ and then defecting is a best response against itself. In case $x_i \geq \underline{x}$ for $i = 1, 2$, each player is either a conditional or an unconditional cooperator. With two conditional cooperators three equilibria exist in the second stage subgame. But the same reasoning as for the Leap mechanism implies that only $C-C$ can occur on the equilibrium path of the entire game. If at least one of the players is an unconditional cooperator, this is immediate, as then the second stage subgame has only $C-C$ as equilibrium outcome. Hence $x_i^* \geq \underline{x}$ and $\delta_i(x_i^*, x_j^*) = 0$ (for $i = 1, 2$) are also equilibrium outcomes. The proposition divides this range into equilibria where both players are conditional cooperators (part (ii)), one is a conditional cooperator and the other one an unconditional cooperator (part (iii)), and equilibria in which both players are unconditional cooperators (part (iv)). □

Proof of proposition 2:

We proceed in three steps:

(i) Any strategy leading to dropping out first (strictly) between 0 and $\underline{x}$ is (iteratively) weakly dominated.

To show this, suppose player i drops out first somewhere between 0 and $\underline{x}$.

In this case playing C in the ensuing second stage game is strictly dominated for player i. Realizing that player i will defect for sure, player j earns the most by dropping out immediately after player i and defect as well. All other responses of player j are therefore (iteratively) weakly dominated. Dropping out first between 0 and $\underline{x}$ thus leads to outcome $D-D$ and punishment with probability $\beta > 0$. Dropping out immediately at zero and playing D then yields player i always weakly more.

(ii) Any strategy leading to dropping out between $\underline{x}$ and $\overline{x}$ is (iteratively) weakly dominated.

First, suppose that player j drops out below $\underline{x}$. Then player i cannot do better than dropping out immediately as well (see (i) above). Next suppose that player j drops out first somewhere between $\underline{x}$ and $\overline{x}$. Then he turns himself into a conditional cooperator. If player i also drops out before $\overline{x}$, she becomes a conditional cooperator as well. The second stage subgame then allows three equilibria ($C-C$, $D-D$ and a mixed one) and the maximum payoff that player i can attain equals c. If instead of dropping out in the conditional cooperator range, player i chooses a possible punishment level above $\overline{x}$, she turns herself into an unconditional cooperator, as choosing D in the second stage game is then strictly dominated for her. Realizing this, player j will choose C in response (because choosing D is strictly dominated for player j). Staying in the mechanism until at least $\overline{x}$ thus yields player i a payoff of c for sure and thus weakly more than dropping out in the conditional cooperator range. A similar argument applies regarding dropping out first between $\underline{x}$ and $\overline{x}$; staying in until at least $\overline{x}$ is weakly better as this leads to outcome $C-C$ (and payoff c) for sure.

(iii) Any strategy leading to dropping out first at zero is (iteratively) weakly dominated.

If player i drops out first at zero, she has a dominant strategy to play D in the second stage game. Realizing this, player j earns the most by dropping out immediately after player i and defect as well (i.e. all other responses of player j are (iteratively) weakly dominated). Dropping out first at zero thus yields a payoff equal to d. Next suppose player i does not drop out first at zero. If the other player does so, player i's best response is to do so as well and outcome $D-D$ results. This gives player i the same payoffs as dropping out first. If the other player does not drop out at zero, from *(i)* and *(ii)* above it is then weakly dominated for both players to drop out before $\overline{x}$. Therefore, both will choose a possible punishment level above $\overline{x}$ and outcome $C-C$ results. This gives both players a payoff equal to $c > d$.

From *(i)* through *(iii)* it follows that dropping out first below $\overline{x}$ is weakly dominated. □

For Online Publication: A.2 The case $t - c > d - s$

For expositional reasons we focused in the main text on the case where the incentive to defect is weaker if the other cooperates than if the other defects, i.e. where $t - c \leq d - s$. In this appendix we show that qualitatively the same

predictions are obtained for the opposite case where $t-c > d-s$. Also then the gradual mechanism is predicted to perform better in fostering cooperation.

We first present the equivalent of Lemma 1 for the general case in which no assumptions are made on how $t-c$ compares to $d-s$.

Lemma 2 *If the other player j has chosen $x_j > 0$ in the first stage, the equilibrium strategy of player i (with $i \neq j$) in the second stage is as follows:*

(i) If $x_i < \min\left\{t-c, \frac{d-s+(1-\beta)p}{\beta}\right\} \equiv \underline{x}'$, then player i will be a defector, that is, $\delta_i^*(x_i, x_j) = 1$ for all $x_j > 0$;

(ii) If $\underline{x}' \leq x_i \leq \max\left\{t-c, \frac{d-s+(1-\beta)p}{\beta}\right\} \equiv \overline{x}'$, then:

(a) if $t-c \leq \frac{d-s+(1-\beta)p}{\beta}$, player i will be a conditional cooperator; that is, player i prefers to cooperate if the other player cooperates and prefers to defect if the other player defects;

(b) if $t-c > \frac{d-s+(1-\beta)p}{\beta}$, player i will be a reverse cooperator; that is, player i prefers to cooperate if the other player defects and prefers to defect if the other player cooperates;

(iii) If $x_i > \overline{x}'$, then player i will be an unconditional cooperator, that is, $\delta_i^*(x_i, x_j) = 0$ for all $x_j > 0$.

If the other player j has chosen $x_j = 0$ in the first stage, then for player i the same strategies hold with $\underline{x}' = \min\left\{t-c, \frac{d-s}{\beta}\right\}$ and $\overline{x}' = \max\left\{t-c, \frac{d-s}{\beta}\right\}$.

Proof of Lemma 2:

Cases (i), (iia) and (iii) are analogous to cases (i), (ii) and (iii) in Lemma 1. For the remaining case (iib), note that $x_i \leq t-c$ implies that player i prefers to choose D if the other chooses C while $\frac{d-s+(1-\beta)p}{\beta} \leq x_i$ implies that i prefers C if the other chooses D. □

A first observation is that the analysis in the main text remains valid as long as part (iib) of Lemma 2 does not apply. The new part (iib) opens up the possibility that a player turns himself into a "reverse cooperator" if he chooses a punishment level in the intermediate range. A reverse cooperator prefers to do the exact opposite of what the other player does.

In the remainder we focus on the case not covered by the main text where $t-c > \frac{d-s+(1-\beta)p}{\beta}$. This corresponds to $\beta > \frac{d-s+p}{t-c+p}$.[25] Players may then turn themselves into reverse cooperators. It can be easily seen that under the Leap mechanism they will never do so. This follows because the best response to the other player being a reverse cooperator is to choose $x_i = 0$ oneself (yielding the

[25] For the case $t-c > d-s$ considered in this appendix, there always exist values of β such that $\frac{d-s+p}{t-c+p} < \beta \leq 1$. For $\beta \leq \frac{d-s+p}{t-c+p}$ the propositions presented in the main text apply.

maximum payoff of t). This induces the other player to choose $x_j = 0$ as well. Hence no equilibria with reverse cooperators exist.

Things are different under the Gradual mechanism. Here equilibria in which both players are reverse cooperators do exist. The intuition runs as follows. If both players are reverse cooperators, the (collapsed) second stage subgame has three equilibria: $C-D$, $D-C$ and a mixed strategy equilibrium. Coordinating on one of the pure equilibria $C-D$ or $D-C$ cannot occur on the equilibrium path of the entire game, as the player supposed to cooperate earns only $s-p$ and would have been better off dropping out immediately at zero and defect, yielding d. Yet coordinating on the mixed equilibrium can occur. This holds because, unlike for the case in which the other player is a conditional cooperator (case (iia)), if the other is a reverse cooperator there is no incentive to become an unconditional cooperator oneself. This would namely lead to the disadvantageous $C-D$ outcome. Therefore, mixing is possible on the equilibrium path of the entire game (only) if expected payoffs exceed d, the minimum payoff a player can secure himself by dropping out immediately at zero. Earning at least d requires that in the mixed equilibrium, the other player j defects with sufficiently low probability. As $\frac{\partial \delta_j}{\partial x_i} < 0$, this in turn requires that player i himself chooses a high enough punishment level $x_i \geq \widehat{x}$. Threshold $\widehat{x}$ is determined such that player i at least earns d in equilibrium. The following proposition makes the above intuitive predictions precise.

Proposition 3 *For all $\beta > \beta^* \equiv \frac{d-s+p}{t-c+p}$ there is a unique subgame-perfect equilibrium outcome in the Leap mechanism:*

(i) $x_i^* = 0$, *and* $\delta_i^*(0,0) = 1$ *for* $i = 1, 2$;

In the Gradual mechanism two additional sets of equilibrium outcomes exist:

(ii) $\widehat{x} \equiv \frac{(t-d)(d-s+p)-\beta p(c-d)}{(d-s+p)+\beta(c-d)} \leq x_i^* \leq t-c = \overline{x}'$,
and $\delta_i^*(x_i^*, x_j^*) = \frac{(t-x_j^*)-c}{(t-x_j^*)-c+(s-p)-[d-\beta(p+x_j^*)]}$ *for* $i = 1, 2$ *(and* $i \neq j$*);*

(iii) $x_i^* > \overline{x}'$, *and* $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*).*

Proof of Proposition 3:

(a) Leap mechanism. From the proof of Proposition 1 it follows that there are no equilibria in which $0 < x_i < \underline{x}'$ or $x_i > \overline{x}'$ (just replace $\underline{x}$ by $\underline{x}'$ and $\overline{x}$ by $\overline{x}'$ in the proof of Proposition 1) and that $x_1 = x_2 = 0$ followed by mutual defection can be supported as equilibrium. It remains to be shown that if $\beta > \beta^*$, no equilibria with $\underline{x}' \leq x_i \leq \overline{x}'$ exist. For $\beta > \beta^*$ this range equals $\frac{d-s+(1-\beta)p}{\beta} \leq x_i \leq t-c$. In that case player i is a reverse cooperator. The best response of the other player j is then $x_j = 0$, yielding her the maximum possible payoff of t. But against $x_j = 0$, choosing $\frac{d-s+(1-\beta)p}{\beta} \leq x_i \leq t-c$ is not a best response, as choosing $x_i = 0$ would yield player i more.

(b) Gradual mechanism. Again it follows from the proof of Proposition 1 that there are no equilibria in which $0 < x_i < \underline{x}'$ and that $x_1 = x_2 = 0$

followed by mutual defection can be supported as equilibrium. So consider the case in which $x_i \geq \underline{x}'$ for $i = 1, 2$. Each player is then either a reverse or an unconditional cooperator. With two reverse cooperators three equilibria exist in the second stage subgame: $C - D$, $D - C$ and a mixed one. On the equilibrium path of the entire game $C - D$ and $D - C$ cannot occur; the player supposed to cooperate would then earn $s - p$, less than the amount d the player could at least get by dropping out immediately at $x_i = 0$. The mixed equilibrium may potentially occur though, because it can yield both players more than d.

In the mixed equilibrium the mixing probability δ_i of player i should make the other player j indifferent between C and D, i.e.: $(1 - \delta_i) \cdot c + \delta_i \cdot (s - p) = (1 - \delta_i) \cdot (t - x_j) + \delta_i \cdot (d - \beta p - \beta x_j)$. Rewriting this yields the expression for $\delta_i^*(x_i^*, x_j^*)$ in part (ii) of the proposition. Differentiating we obtain that $\frac{\partial \delta_i^*}{\partial x_j} < 0$ for $\beta > \beta^*$. A necessary condition for mixing to occur on the equilibrium path is that players at least earn d. The expected payoff of player j can be calculated from $(1 - \delta_i^*) \cdot c + \delta_i^* \cdot (s - p)$. For this to exceed d it must hold that $\delta_i^* \leq \frac{c-d}{c-s+p}$. Plugging in the expression for δ_i^* and rewriting yields that $x_j^* \geq \frac{(t-d)(d-s+p)-\beta p(c-d)}{(d-s+p)+\beta(c-d)} \equiv \widehat{x}$ is required (it holds that $\underline{x}' \leq \widehat{x} \leq \overline{x}'$). To support $\widehat{x} \leq x_i^*, x_j^* \leq \overline{x}'$ and subsequent mixing, off path behavior is governed by Lemma 2 for deviations outside $[\underline{x}', \overline{x}']$. After deviations $x_i \neq x_i^*$ within $[\underline{x}', \overline{x}']$, the for player i disadvantageous pure equilibrium follows (i.e. player i choosing C, the other player D).

One reverse cooperator together with an unconditional cooperator would lead to the $D - C$ outcome, which cannot happen on the equilibrium path (as the cooperating player gets less than d). Two unconditional cooperators can be supported on the equilibrium path by assuming that in the reverse cooperator range $[\underline{x}', \overline{x}']$ players coordinate on the mixed equilibrium of the second stage subgame. Expected payoffs in this mixed equilibrium are below c, so neither player has an incentive to deviate in that range. □

The equilibria of parts (i) and (iii) of Proposition 3 are as before, with only defectors and unconditional cooperators respectively. In the equilibria of part (ii) both players have turned themselves into reverse cooperators in the first stage. They subsequently use mixed strategies in the second stage subgame, with the probability of coordinating on $C-C$ (equal to $(1-\delta_1^*)(1-\delta_2^*)$) increasing in the punishment levels chosen in the first stage.

Overall we conclude that when the incentive to defect is stronger if the other player cooperates then when the other defects, coordination on a cooparative outcome can only be sustained in the Leap mechanism if β is sufficiently low (i.e. $\beta \leq \beta^*$ is needed). In contrast, under the Gradual mechanism cooperation continues to be an equilibrium outcome for all values of $\beta \in (0, 1]$. The Gradual mechanism is therefore predicted to perform better also when $t-c > d-s$. Note though that the case for cooperation under the Gradual mechanism then also becomes less focal, as an equivalent of Proposition 2 does not hold for $\beta > \beta^*$.

For Online Publication: A.3 - Inequity aversion

In this Appendix we report the equilibrium analysis of our three stage game for inequity averse preferences in the spirit of Fehr and Schmidt (1999). For ease of exposition we assume that players dislike being behind, but do not dislike being ahead. Furthermore, we also assume that players are homogeneous. All players' preferences are thus given by $U_i(y_i, y_j) = y_i - \alpha \max\{y_j - y_i, 0\}$, where y_i and y_j denote the monetary payoffs players get and $\alpha > 0$ a parameter measuring a player's level of inferiority aversion.

For these preferences we show the following results. With moderate levels of inequity aversion (i.e. α taking an intermediate value) the same types of equilibria exist as in the main analysis. Lemma 1 and Proposition 1 in the main text then remain qualitatively valid, with the thresholds separating the different cases now also depending on the level of inequity aversion α. For low levels of inequity aversion the two mechanisms do not differ in their equilibrium predictions. Another important implication is that players' types become interdependent. With inequity averse preferences a player becomes an unconditional cooperator only if he himself chooses a high possible punishment level ánd the other player chooses a possible punishment level that is neither very high nor very low.[26]

The formal derivation of the subgame-perfect equilibria itself is straightforward but tedious and – apart from the main results just summarized – not particularly insightful. We start solving the game backwards as before. Unlike the analysis in Section 3.2, however, here we do not make a priori assumptions about punishment behavior of the players and therefore start the equilibrium analysis with the third punishment stage. Lemma 3 gives players' equilibrium punishment behavior. Here N and P denote no punishment and punishment, respectively.

Lemma 3 *Suppose that in the first stage players choose possible punishment levels x_i and x_j, respectively. Then depending on the outcome of the second stage prisoner's dilemma, the equilibrium outcomes in the third stage are as follows:*

(i) After outcomes $C - C$ and $D - D$:

(a) If $\min\{x_i, x_j\} < \left(\frac{1}{\alpha} + 1\right) p \equiv p(\alpha) : N - N$

(b) If $\min\{x_i, x_j\} \geq p(\alpha) : N - N$, $P - P$, mixed equilibrium

[26] The intuition here is that, if the other player j chooses a high x_j, he will also punish when he defects and player i cooperates, i.e. after outcome $C - D$, in order to counterbalance the punishment done by the cooperating player i. By choosing C in response to D player i then cannot avoid punishment, so he is better off choosing D. Similarly, if x_j is very low, player i is never willing to punish, because the decrease in the other player's payoffs (x_j) does not outweigh the costs of punishment (p). If player i is not willing to punish, neither is player j when a symmetric outcome in the prisoner's dilemma is reached. After outcome $D-D$ players thus never punish, making D a best response to the other choosing D. Overall, therefore, if x_j of the other player is either very low or very high, player i can be a conditional cooperator at most.

(ii) After outcomes $C-D$, with x_C the possible punishment level of the cooperator and x_D of the defector, and $\underline{\alpha} \equiv \left(\frac{p}{s-t}\right)$:

Table 7: Equilibria after $C-D$ for the different possible punishment levels

	$x_D < p(\alpha)$	$p(\alpha) \leq x_D \leq t-s+p(\alpha)$	$x_D > t-s+p(\alpha)$
$x_C < p(\alpha)$	$N-N$	$\alpha < \underline{\alpha} : N-N$ $\alpha \geq \underline{\alpha} : P-N$	$\alpha < \underline{\alpha} : N-N$ $\alpha \geq \underline{\alpha} : P-N$
$x_C \geq p(\alpha)$	$N-N$	$\alpha < \underline{\alpha} : N-N$ $\alpha \geq \underline{\alpha} : P-N$	$\alpha < \underline{\alpha} : N-N$, $P-P$, mixed $\alpha \geq \underline{\alpha} : P-P$

Proof of Lemma 3:

(i) Table 8 shows players' utility for given combinations of punishment choices after outcome $C-C$. From $c > c-p-\alpha \max\{p-x_j, 0\}$ it immediately follows that if the other player does not punish, it is a best response not to punish as well. Thus $N-N$ is always an equilibrium. Choosing N in response to P is a best response for player i whenever $-\alpha \max\{x_i-p, 0\} > -p-\alpha \max\{x_i-x_j, 0\}$. (Throughout the analysis ties are broken arbitrarily.) This inequality is equivalent to $\min\{x_1, x_2\} < p(\alpha)$. Hence for $\min\{x_1, x_2\} < p(\alpha)$ players have a dominant strategy not to punish. In case $\min\{x_1, x_2\} \geq p(\alpha)$, choosing P is a best response against P. The punishment subgame then has three Nash equilibria: $N-N$, $P-P$, and a mixed strategy equilibrium. In the latter players punish with probability $\beta(x_i, x_j) \equiv \frac{p}{\alpha(\min\{x_i, x_j\}-p)}$. The reasoning for outcome $D-D$ is similar, just replace c by d in Table 8.

Table 8: Players' utility in the punishment stage after outcome $C-C$

	N	P
N	c, c	$c-x_1-\alpha \max\{x_1-p, 0\}$, $c-p-\alpha \max\{p-x_1, 0\}$
P	$c-p-\alpha \max\{p-x_2, 0\}$ $c-x_2-\alpha \max\{x_2-p, 0\}$	$c-x_1-p-\alpha \max\{x_1-x_2, 0\}$, $c-x_2-p-\alpha \max\{x_2-x_1, 0\}$

(ii) Table 9 shows players' utility for given combinations of punishment choices after outcome $C-D$. Here without loss of generality we assume that (row) player 1 cooperates and (column) player 2 defects.

First consider the deviating player 2. He always prefers to choose N in response to N. If the other, cooperating player 1 chooses P, choosing N is a best response for player 2 whenever $-\alpha \max\{s-p-t+x_2, 0\} \geq -p-\alpha \max\{s-x_1-t+x_2, 0\}$. This condition is certainly satisfied when $-\alpha(s-p-t+x_2) \geq -p$, i.e. for $x_2 \leq t-s+p(\alpha)$. Similarly so, the condition is certainly satisfied whenever $-\alpha(x_2-p) > -p-\alpha(x_2-x_1)$. This reduces to $x_1 < p(\alpha)$. So, if $x_1 < p(\alpha)$ or $x_2 \leq t-s+p(\alpha)$, the deviating player 2 necessarily chooses N. For these cases the cooperating player 1 prefers to choose N as well if

Table 9: Players' utility in the punishment stage after outcome $C - D$

	N	P
N	$s - \alpha(t - s)$, t	$s - x_1 - \alpha \max\{t - p - s + x_1, 0\}$, $t - p - \alpha \max\{s - x_1 - t + p, 0\}$
P	$s - p - \alpha \max\{t - x_2 - s + p, 0\}$ $t - x_2 - \alpha \max\{s - p - t + x_2, 0\}$	$s - p - x_1 - \alpha \max\{t - x_2 - s + x_1, 0\}$, $t - p - x_2 - \alpha \max\{s - x_1 - t + x_2, 0\}$

$-\alpha(t-s) < -p - \alpha \max\{t - x_2 - s + p, 0\}$. This condition holds whenever either $\alpha < \left(\frac{p}{s-t}\right) \equiv \underline{\alpha}$ or $\alpha \geq \underline{\alpha}$ and $x_2 < p(\alpha)$. This yields all cases in Table 7 except for the bottom right cell. For that case, player's 2 best response to P is to punish as well. If $\alpha \geq \underline{\alpha}$ player 1 has a dominant strategy to punish and only $P - P$ is possible. Otherwise, player 1 prefers to match player 2's choice and three types of equilibria result ($N - N$, $P - P$, and a mixed equilibrium). □

To better understand the lemma, suppose that player 1 is behind in monetary payoffs. By spending p on punishment he can reduce inequality by $x_2 - p$ (assuming that player 1 is still behind after punishment). Player 1 is only willing to do so if the monetary costs of punishment p fall short of the benefits $\alpha\,(x_2 - p)$, viz. if $x_2 \geq p(\alpha)$. Intuitively, punishment should be sufficiently effective for player 1 to be willing to carry it out. Note that $p(\alpha)$ is decreasing in α, with $p(\alpha) \to \infty$ for $\alpha \to 0$ and $p(\alpha) \to p$ for $\alpha \to \infty$. Selfish players are never willing to punish, while very inequity averse players are willing to punish if the monetary harm imposed is larger than the costs borne.

Lemma 3 reveals that there may exist multiple equilibria in the punishment stage. To facilitate the equilibrium analysis of the entire game, we make assumptions such that we can focus on one outcome in these cases. First, after $C - C$ there are multiple equilibria if the possible punishment levels are sufficiently high. Because there is no compelling reason to punish after cooperation, however, equilibrium $N - N$ serves as obvious focal point. In the ensuing analysis we therefore assume that players coordinate on the equilibrium in which they never punish after outcome $C - C$. Second, the situation after outcome $D - D$ is less clear. Since the other player defected, each player has reasons to punish the other for not cooperating. At the same time, each player defected himself as well and thus behaved equally "badly" as the other player did. Therefore, both $N - N$ and $P - P$, together with the mixed equilibrium, seem plausible; there is no focal point among the three equilibria.[27] To simplify the equilibrium analysis while still being able to capture the difference between the three different cases, we assume that players punish with probability $\beta \in [0, 1]$ after the $D - D$ outcome if there are multiple equilibria in the punishment stage.[28]

[27] In the mixed strategy equilibrium players punish with probability $\beta(x_i, x_j) \equiv \frac{p}{\alpha(\min\{x_i, x_j\} - p)}$. Note that $\beta(x_i, x_j) = \beta(x_j, x_i)$, so also the mixed strategy equilibrium is fully symmetric in the players' punishment strategies.

[28] This approach seems more 'reduced form' than it actually is. By setting β equal to

Third, we assume that $\alpha \geq \underline{\alpha}$, i.e. that players sufficiently dislike being behind. This simplifies the equilibrium analysis in two ways: (i) after outcome $C-D$ then always a single punishment equilibrium exists, even when the defector chose a very high possible punishment level $x_D > t-s+p(\alpha)$, and (ii) it ensures that the cooperator always punishes after outcome $C-D$, if punishment is sufficiently effective (i.e., $x_D \geq p(\alpha)$). Given that p is small and $t-s$ is the largest payoff difference in the prisoner's dilemma, assuming $\alpha \geq \underline{\alpha}$ seems very plausible. For the parameters in the experiment it reduces to $\alpha > 0.06$. Finally, we assume that α cannot be too large as well, i.e. $\alpha \leq \frac{1}{\beta}$. This assumption ensures that if after outcome $C-D$ the defecting player punishes as well (just like the cooperating player), a player always prefers to choose D in response to D. Intuitively this makes sense: if punishment cannot be avoided by choosing C, it is better to defect as well.

Assumption A1 summarizes the above assumptions.

Assumption A1 In the following we assume that:

(i) after outcome $C-C$, if $\min\{x_i, x_j\} \geq p(\alpha) \equiv \left(\frac{1}{\alpha}+1\right)p$ players coordinate on equilibrium $N-N$;

(ii) after outcome $D-D$, if $\min\{x_i, x_j\} \geq p(\alpha)$ players punish with probability $\beta \in [0,1]$;

(iii) $\alpha \geq \underline{\alpha} \equiv \left(\frac{p}{t-s}\right)$;

(iv) $\alpha \leq \frac{1}{\beta}$.

Using these assumptions, Lemma 4 below describes how players behave in the stage two prisoner's dilemma game as a function of the possible punishment levels. This lemma is the equivalent of Lemma 1 in the main text and the more general Lemma 2 in Appendix A.2.[29]

Lemma 4 *Suppose players behave in the third stage according to Lemma 3 and also suppose that the conditions of Assumption A1 hold. If the other player j has chosen x_j in the first stage, the equilibrium strategy of player i (with $i \neq j$) in the second stage is as follows:*

---

one of the three equilibrium values in the analysis below (i.e., either $\beta = 0$, $\beta = 1$ or $\beta = \beta(x_i, x_j)$), a true subgame perfect equilibrium of the overall game results. For instance, for $\beta = \beta(x_i, x_j)$ the expression for $X(x_i, x_j)$ in Lemma 4 below reduces to $X(x_i, x_j) \equiv \frac{d-s+(1+\beta(x_i,x_j)\alpha)p+\alpha\max\{(t-x_j-s+p,0\}}{\beta(x_i,x_j)(1+\alpha)}$ for $p(\alpha) \leq x_j \leq (t-s)+p(\alpha)$ and $X(x_i, x_j) \equiv \infty$ elsewhere. For this $X(x_i, x_j)$ Lemma 4 then immediately applies for equilibria in which the mixed strategy punishment equilibrium always follows after $D-D$.

[29] To see the correspondence, note that: (i) $p(\alpha)$ would vanish in Lemma 4 if punishment were automatic (as in the main text) rather than given by equilibrium behavior (as described in Lemma 3), and (ii) for $\alpha = 0$ we have $X(x_i, x_j) = \frac{d-s+(1-\beta)p}{\beta}$. Under (i) and (ii) we obtain $\underline{x}'' = \underline{x}'$ and $\overline{x}'' = \overline{x}'$ and Lemma 4 reduces to Lemma 2 in Appendix A.2.

(i) *If* $x_i < \max\{p(\alpha),\ \min\{t-c, X(x_i,x_j)\}\} \equiv \underline{x}''$, *then player* i *will be a defector, that is,* $\delta_i^*(x_i,x_j)=1$ *for all* $x_j \geq 0$;

(ii) *If* $\underline{x}'' \leq x_i \leq \max\{p(\alpha), t-c, X(x_i,x_j)\} \equiv \overline{x}''$, *then:*

- *(a)* *if* $t-c \leq X(x_i,x_j)$, *player* i *will be a conditional cooperator; that is, player* i *prefers to cooperate if the other player cooperates and prefers to defect if the other player defects;*
- *(b)* *if* $t-c > X(x_i,x_j)$, *player* i *will be a reverse cooperator; that is, player* i *prefers to cooperate if the other player defects and prefers to defect if the other player cooperates;*

(iii) *If* $x_i > \overline{x}''$, *then player* i *will be an unconditional cooperator, that is,* $\delta_i^*(x_i,x_j)=0$ *for all* $x_j > 0$.

Here $X(x_i,x_j) \equiv \frac{d-s+(1-\beta)(1+\beta\alpha)p-\beta^2\alpha\max\{x_i-x_j,0\}+\alpha\max\{(t-x_j-s+p,0\}}{\beta+(1-\beta)\beta\alpha}$ for $p(\alpha) \leq x_j \leq (t-s)+p(\alpha)$ and $X(x_i,x_j) \equiv \infty$ for $x_j < p(\alpha)$ or $x_j > (t-s)+p(\alpha)$.

Proof of Lemma 4:

We first show that if $x_i < p(\alpha)$, player i has a dominant strategy to defect. Let $x_i < p(\alpha)$. Then from Lemma 3 we have that $N-N$ follows after $D-D$. For player i choosing D is thus a best response against D (as choosing C would lead to a lower monetary payoff and possibly disutility from disadvantageous inequality). After outcome $D-C$ (with player i choosing D), we again have from Lemma 3 that $N-N$ follows (cf. $x_D < p(\alpha)$ in Table 7). Given no punishment, choosing D is indeed a best response for player i against C.

Next consider the case $x_i \geq p(\alpha)$. Here four different situations have to be considered. First, suppose $x_j < p(\alpha)$. W.l.o.g., let $x_1 \geq p(\alpha)$ and $x_2 < p(\alpha)$. Table 10 then reflects the players' utility in the merged second and third stage. It immediately follows that player 1 always prefers to choose D against D. If player 2 chooses C, player 1 prefers D iff $t-x_1-\alpha\max\{s-p-t+x_1,0\} \geq c$. The l.h.s. is strictly decreasing in x_1 and equals c for $x_1 = t-c$. Hence, for $x_1 < t-c$ player 1 is a defector, for $x_1 \geq t-c$ a conditional cooperator.

Table 10: Players' utility in the PD-game, with $x_1 \geq p(\alpha)$ and $x_2 < p(\alpha)$

	C	D
C	c, c	$s-\alpha(t-s)$, t
D	$t-x_1-\alpha\max\{s-p-t+x_1,0\}$, $s-p-\alpha\max\{t-x_1-s+p,0\}$	d, d

Second, suppose $p(\alpha) \leq x_i, x_j \leq (t-s)+p(\alpha)$. Collapsing the third and second stage, Table 11 reflects the payoffs for this case. Again player 1 prefers to choose D against C iff $x_1 \leq t-c$. If player 2 chooses D, player

Table 11: Players' utility in the PD-game, with $p(\alpha) \leq x_1, x_2 \leq (t-s) + p(\alpha)$

	C	D
C	c,c	$s - p - \alpha \max\{t - x_2 - s + p, 0\}$, $t - x_2 - \alpha \max\{s - p - t + x_2, 0\}$
D	$t - x_1 - \alpha \max\{s - p - t + x_1, 0\}$, $s - p - \alpha \max\{t - x_1 - s + p, 0\}$	$d - \beta(x_1 + p) - (1-\beta)\,\beta\alpha\,(x_1 - p) - \beta^2\alpha \max\{x_1 - x_2, 0\})$, $d - \beta(x_2 + p) - (1-\beta)\,\beta\alpha\,(x_2 - p) - \beta^2\alpha \max\{x_2 - x_1, 0\})$

1 prefers D whenever $d - s + (1-\beta)(1+\beta\alpha)p + \alpha \max\{(t - x_2 - s + p, 0\} \geq (\beta + (1-\beta)\beta\alpha)\, x_1 + \beta^2\alpha \max\{x_1 - x_2, 0\})$, i.e. whenever $x_1 \leq X(x_1, x_2)$.

Third, if $x_i > (t-s) + p(\alpha)$ and $p(\alpha) \leq x_j \leq (t-s) + p(\alpha)$, payoffs are as in Table 11, with the single exception that after outcome $D-C$ the defecting player i gets $t - x_i - p - \alpha \max\{s - x_j - t + x_i, 0\}$ instead of $t - x_i - \alpha \max\{s - p - t + x_i, 0\}$, because the defector now also punishes. For $x_i \geq t - c - p$ (and thus for all $x_i > (t-s) + p(\alpha)$) player i then prefers C against C; player i is thus either a conditional or an unconditional cooperator. The former case applies if $x_i \leq X(x_i, x_j)$, the latter if $x_i > X(x_i, x_j)$.

Finally, let $x_i > p(\alpha)$ and $x_j > (t-s) + p(\alpha)$. In that case, after outcome $C - D$ also the defecting player punishes. Choosing C in response to D thus yields player i a payoff of $s - x_i - p - \alpha \max\{t - x_j - s + x_i, 0\}$, while choosing D gives $d - \beta(x_i + p) - (1-\beta)\,\beta\alpha\,(x_i - p) - \beta^2\alpha \max\{x_i - x_j, 0\})$. Therefore D is preferred whenever $d - s + \alpha \max\{t - x_j - s + x_i, 0\} - \beta^2\alpha \max\{x_i - x_j, 0\}) \geq (1-\beta)\,(\beta\alpha - 1)\, x_i - (1-\beta)\,(\beta\alpha + 1)\, p$. The l.h.s. is necessarily positive. A sufficient condition for the r.h.s. to be negative is $\alpha \leq \frac{1}{\beta}$ (cf. Assumption A1).

Taken all cases together it follows that, whenever $x_i < p(\alpha)$, or when $x_i \geq p(\alpha)$ and $x_i < \min\{t - c, X(x_i, x_j)\}$, player i is a defector. This gives part (i) of the lemma. If $x_i > p(\alpha)$ and $x_i > \min\{t - c, X(x_i, x_j)\}$, player i prefers to choose C against C (if $x_i > t - c$) or C against D (if $x_i > X(x_i, x_j)$), or both. Parts (iia), (iib) and (iii) give the three possible cases. □

Lemma 4 has the same structure as Lemma 1 in the main text. Yet, there are two important differences. First, with inequity averse preferences players may turn themselves into reverse cooperators for some parameter constellations (even when $t - c \leq d - s$). Reverse cooperators prefer to do the exact opposite of what the other player does in the prisoner's dilemma. In Appendix A.2 we already discussed that this does not affect our qualitative predictions.

Second, and more importantly, types are no longer independent. In particular, the threshold for player i to turn himself into an unconditional cooperator not only depends on his own possible punishment level x_i, but also on the possible punishment level of the other player x_j. In fact, if $x_j < p(\alpha)$ or $x_j > (t-s) + p(\alpha)$, then $X(x_i, x_j) \equiv \infty$ and $x_i < \overline{x}''$ necessarily. In that case player i can only either be a defector (for $x_i \leq \max\{p(\alpha),\ t - c\}$) or a conditional cooperator (for $x_i > \max\{p(\alpha),\ t - c\}$). The intuition here is that for $x_j < p(\alpha)$, player i is never willing to punish player j. Realizing this, player j does not punish as well after outcome $D - D$. Given no punishment after

$D - D$, players strictly prefer D over C in response to D and thus cannot turn themselves into an unconditional cooperator. In a similar vein, in case $x_j > (t-s) + p(\alpha)$, player j punishes player i if he himself is the single defector (i.e. after outcome $C - D$, with i choosing C). Player j does so to reduce inequality, because he is punished himself by player i. For player i it is then not a best response to choose C in response to D, hence i cannot become an unconditional cooperator.

The possibility of reverse cooperators may only occur for high values of α. To avoid distinguishing many cases we focus on intermediate levels of inferiority aversion for which this cannot happen. (Moreover, the intuition behind the reverse cooperator case has already been discussed in Appendix A.2.) Also the type interdependency may potentially lead to many different cases, as mixed strategy equilibria in the prisoner's dilemma game when both players are conditional cooperators cannot immediately be ruled out.[30] We therefore simply assume that players do not coordinate on the mixed equilibrium if both of them happen to be conditional cooperators. We thus make the following additional assumptions:[31]

Assumption A2 In the following we assume that:

(i) $\alpha \leq \left(\frac{(d-s)-(t-c)+p}{t-c-p} \right)$;

(ii) If, for the chosen possible punishment levels x_i (for $i = 1, 2$) multiple equilibria exist in the second stage prisoner's dilemma, players do not coordinate on the mixed equilibrium, but either on $C - C$ or $N - N$.

Part (i) of Assumption A2 ensures that case (iib) in Lemma 4 cannot occur.[32] Using Lemma 4 we can finally establish the equivalent of Proposition 1.

Proposition 4 *Suppose Assumptions A1 and A2 hold. In both versions of the mechanism there are multiple subgame-perfect equilibrium outcomes:*

(a) *if* $t - c \geq p(\alpha) \equiv \left(\frac{1}{\alpha} + 1\right) p$*:*

(i) $x_i^* < p(\alpha)$, $\delta_i^*(x_i^*, x_j^*) = 1$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment;*[33]

[30] If $x_i > (t-s) + p(\alpha)$ for $i = 1, 2$ both players are conditional cooperators and cannot turn themselves into unconditional cooperators by choosing a different possible punishment level (taking the punishment level of the other as given).

[31] Part (i) of Assumption A2 is not automatically satisfied given our assumption (iv) in Assumption A1, nor vice versa. Effectively we thus assume $\alpha \leq \overline{\alpha} \equiv \min \left\{ \left(\frac{(d-s)-(t-c)+p}{t-c-p} \right), \frac{1}{\beta} \right\}$

[32] This follows from observing that for $x_i < t - c$ and $x_j > p(\alpha)$ it holds that $X(x_i, x_j) \geq \frac{d-s+(1-\beta)(1+\beta\alpha)p-\beta^2\alpha(t-c-p(\alpha))}{\beta+(1-\beta)\beta\alpha}$. The latter exceeds $t - c$ whenever $\alpha \leq \frac{d-s-\beta(t-c)+(1-\beta+\beta^2)p}{\beta(t-c-p)}$. The r.h.s. is decreasing in β and equals $\frac{(d-s)-(t-c)+p}{t-c-p}$ for $\beta = 1$.

[33] In fact, here the set of equilibria is actually larger. The necessary requirement is that $x_i^* < p(\alpha)$ for at least one of the two players, see the proof of the proposition for a full characterization. All these equilibria result in outcome $D - D$ without punishment.

(ii) $t-c \leq x_i^* \leq \frac{d-s+(1-\beta)(1+\beta\alpha)p+\beta^2\alpha(t-c)+\alpha(c-s+p)}{\beta(1+\alpha)} \equiv \overline{x}^{FS}$, $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment;*

In the Gradual mechanism two additional sets of equilibrium outcomes exist:

(iii) $t - c \leq x_i^* \leq \overline{x}^{FS}$ *and* $x_j^* > \overline{x}^{FS}$, $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment;*

(iv) $x_i^* > \overline{x}^{FS}$, $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment.*

(b) *if* $t - c < p(\alpha)$:

(i) $x_i^* < p(\alpha)$, $\delta_i^*(x_i^*, x_j^*) = 1$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment;*

(ii) $x_i^* \geq p(\alpha)$, $\delta_i^*(x_i^*, x_j^*) = 0$ *for* $i = 1, 2$ *(and* $i \neq j$*) without punishment.*

Proof of Proposition 4:

We consider the two versions of the mechanism separately.

(I) Leap mechanism. We first show that $x_i^* < p(\alpha)$ with $\delta_i^*(x_i^*, x_j^*) = 1$ for $i = 1, 2$ (and $i \neq j$) without punishment always constitutes an equilibrium. Suppose $x_i^* < p(\alpha)$ for $i = 1, 2$. From Lemma 4 it follows that both players are defectors while from Lemma 3 it follows that players do not punish after $D - D$ (given $\min\{x_i, x_j\} < p(\alpha)$). Both players thus obtain d. Choosing another possible punishment level does not change the other player's type (the other player remains a defector) and thus cannot yield more than d. Choosing $x_i < p(\alpha)$ is therefore a best response against the other player choosing $x_j < p(\alpha)$. This gives parts (a-i) and (b-i) for the Leap mechanism.

(In fact, note that $x_i \geq p(\alpha)$ is a best response to $x_j < p(\alpha)$ as well; in that case we have $X(x_i, x_j) = \infty$ and thus any x_i would be a best response. But for $x_i \geq p(\alpha)$ the other player j may want to deviate from $x_j < p(\alpha)$. If $x_i \in [\max\{p(\alpha), t-c\}, (t-s)+p(\alpha)]$, player j could profitably deviate to a high value of x_j such that $x_j \geq X(x_j, x_i)$ is satisfied and j becomes an unconditional cooperator. This would lead to a payoff of c for both. If $x_i > (t-s)+p(\alpha)$ player j cannot turn himself into an unconditional cooperator. A profitable deviation for player j then exists only if, by deviating, player j turns the other player i into an unconditional cooperator and becomes a conditional cooperator himself. This requires for deviation x_j' that $x_j' \leq p(\alpha) + (t-s)$ and $x_i \geq X(x_i, x_j')$. This in turn requires that $x_i\beta(1+\alpha) \geq d - s + (1-\beta)(1+\beta\alpha)p + \beta^2\alpha x_j' + \alpha\max\{(t - x_j' - s + p, 0\}$. The r.h.s. is smallest for $x_j' = t-s+p$. Therefore, for $x_i \in \left[(t-s)+p(\alpha), \frac{d-s+(1-\beta+\beta\alpha)p+\beta^2\alpha(t-s)}{\beta(1+\alpha)}\right]$ no profitable deviation exists. Hence the complete set of equilibria equals: $x_i^* < p(\alpha)$ and $x_j^* < t - c$ or $x_j^* \in \left[(t-s)+p(\alpha), \frac{d-s+(1-\beta+\beta\alpha)p+\beta^2\alpha(t-s)}{\beta(1+\alpha)}\right]$ for $i = 1, 2$ (and $i \neq j$) without punishment. All these equilibria are essentially equivalent in that they result in

outcome $D-D$ without punishment. For ease of presentation, we do no list all these equilibria in the formulation of the proposition.)

Next consider equilibria in which cooperation occurs with positive probability on the equilibrium path. From Lemma 4 this necessarily requires $x_i \geq p(\alpha)$ for $i = 1, 2$. From Assumption A1 we have that players coordinate on $N-N$ after $C-C$, punish with probability β after $D-D$, while after $C-D$ the cooperator certainly punishes. For player i to prefer C against C it must hold that $x_i \geq \max\{p(\alpha), t-c\}$. The former is needed for player j to be willing to carry out punishment, the latter for this punishment to be sufficiently strong to deter player i from choosing D against C if this leads to punishment.

Suppose therefore that $x_i \geq \max\{p(\alpha), t-c\}$ for $i = 1, 2$. Given Lemma 4 and Assumption A2 players are then either conditional or unconditional cooperators. In equilibrium they will necessarily coordinate on the $C-C$ outcome afterwards (note that by Assumption A2 we have excluded coordination on the mixed equilibrium). This holds because, if not, one of the players could profitably deviate to $x = 0$. Both players therefore necessarily earn c on the equilibrium path. A possible deviation from one player might not only change the type of that player, but also the other player's type. Note, however, that a deviation by player i never changes player j into a defector. (Recall that by Assumption A2 we have $t - c \leq X(x_i, x_j)$.) A profitable deviation by player i can only occur if it induces outcome $D-C$ with positive probability. This yields player i at most $t - x_i$ and thus requires that $x_i' \leq t - c$ to be potentially profitable. Now for player j to be willing to choose C in response to D it is required that $x_j > X(x_j, x_i')$ and thus $x_i' \geq p(\alpha)$ necessarily. For $t - c < p(\alpha)$ the two conditions $x_i' \leq t - c$ and $x_i' \geq p(\alpha)$ cannot hold at the same time, thus a profitable deviation does not exist. Both players choosing $x_i \geq \max\{p(\alpha), t-c\} = p(\alpha)$ and coordinating on $C-C$ thus constitutes an equilibrium. This gives part (b-ii) for the Leap mechanism. Next consider the case where $t - c \geq p(\alpha)$. For $x_i' \leq t - c$ (and $x_j \geq \max\{p(\alpha), t-c\}$) it holds that $X(x_j, x_i) = \frac{d-s+(1-\beta)(1+\beta\alpha)p-\beta^2\alpha(x_j-x_i)+\alpha(t-x_i-s+p)}{\beta+(1-\beta)\beta\alpha}$. The r.h.s. is weakly decreasing in x_i. For $x_i = t - c$ the inequality $x_j > X(x_j, x_i)$ can be rewritten as $x_j > \frac{d-s+(1-\beta)(1+\beta\alpha)p+\beta^2\alpha(t-c)+\alpha(c-s+p)}{\beta(1+\alpha)} \equiv \overline{x}^{FS}$. So, if x_j exceeds this threshold, player i can profitably deviate to $x_i' \leq t - c$. If both players choose a possible punishment level between $t - c$ and $\overline{x}^{FS}$, then they do not have an incentive to deviate from these levels, and they are (un)conditional cooperators. This gives part (a-ii) of the proposition.

(II) Gradual mechanism. Parts (a-i) and (b-i) immediately follow from the reasoning under the Leap mechanism. Note that also here the necessary requirement is that $x_i^* < p(\alpha)$ for at least one of the two players; the other possible punishment level can be arbitrarily. Intuitively, if one of the players jumps out before $p(\alpha)$, outcome $D-D$ without punisnment results independent of what the other player does and the possible punishment level at which the other player jumps out is immaterial.

Suppose $x_i \geq \max\{p(\alpha), t-c\}$.for $i = 1, 2$. Then both players are either conditional or unconditional cooperators. By assumption A2 they do not coor-

dinate on the mixed equilibrium, but either on $C-C$ or on $D-D$. On the equilibrium path they necessarily coordinate on $C-C$, for otherwise a deviation to $x=0$ would be profitable. The reasoning above for the Leap mechanism shows that if $t-c<p(\alpha)$ no profitable deviation exists that exploits a (potentially) unconditional cooperator. This gives part (b-ii) for the Gradual mechanism. If $t-c\geq p(\alpha)$ an unconditional cooperator can be exploited by a defector. In that case, since players cannot go back player j is willing to turn himself into an unconditional cooperator only if the other player i at least becomes a conditional cooperator, i.e. for $x_i\geq t-c$. This yields the three remaining cases (a-ii) through (a-iv). □

As discussed in the main text, the Gradual mechanism in general allows for a larger set of equilibria, as players may turn themselves into unconditional cooperators on the equilibrium path. They might be willing to do so once they know that the other player is sufficiently coming along and makes himself vulnerable too (such that he is at least a conditional cooperator). Being an unconditional cooperator cannot occur in the Leap mechanism, as there they would be exploited by the other player by becoming a defector and choose $x_i=0$. Equilibria with unconditional cooperators thus make the difference between the two mechanisms.

Part (a) of Proposition 4 resembles Proposition 1 in the main text. For $t-c\geq p(\alpha)$ the set of equilibria is strictly larger under the Gradual mechanism; in the equilibria of parts (a-iii) and (a-iv) at least one of th players turns himself into an unconditional cooperator by choosing a high possible punishment level. This case applies when α is sufficiently high: $\alpha\geq\frac{p}{t-c-p}$. Taking all assumptions on α into account, for the parameters used in the experiment part (a) of the proposition applies for $\frac{4}{11}\leq\alpha\leq\frac{9}{11}$. Players should thus be at least moderately inferiority averse.

Part (b) of the proposition shows that, if $t-c<p(\alpha)$ (i.e. $\alpha<\frac{4}{11}$), the two versions of the mechanism actually allow the same set of equilibria. The intuition here runs as follows. Suppose player i chooses a high punishment level such that, in principle, he could be an unconditional cooperator. Player j could go along such that players in the end coordinate on the $C-C$ outcome, yielding both of them c. Alternatively, player j could try to exploit player i by choosing a low punishment level and subsequently defect, with the aim of ending up in the $C-D$ outcome. Given that player j is punished after $C-D$, he obtains (at most) $t-x_j$ in monetary terms. Such intended exploitation is thus profitable only if $t-x_j\geq c$, i.e. when $x_j\leq t-c$. But for these possible punishment levels it holds that $x_j<p(\alpha)$, and thus that the other player i cannot be an unconditional cooperator. In short, deviations that might potentially be profitable for player j change the type of the other player i (from unconditional to conditional cooperator), making the deviation unprofitable.

Overall, the analysis in this Appendix shows that with inequity averse players the same type of equilibria exist as in the main analysis; Lemma 1 and Proposition 1 remain qualitatively valid when players are moderately inequity averse. For low levels of inequity aversion the two mechanisms do not differ

in their equilibrium predictions. Another noteworthy feature of the inequity aversion model is that types become interdependent; whether a player becomes an unconditional cooperator not only depends on he himself choosing a high possible punishment level, but also on whether the possible punishment level of the other player is not extreme (i.e. neither very low nor very high).

For Online Publication: Appendix B - Instructions

In this appendix, we present the instructions for the Gradual case, subjects read on the screen at the beginning of the experiment. The instructions for the other two treatments are available upon request from the authors.

INSTRUCTIONS PAGE 1

Welcome to this experiment on decision-making. Please read the following instructions carefully. When everyone has finished reading the instructions and before the experiment starts, you will receive a handout with a summary of the instructions. During the experiment you will be asked to make a number of decisions. Your decisions and the decisions of other participants will determine how much money you earn. At the start of the experiment you will receive a starting capital of 500 points. In addition you will earn money with your decisions. The experiment consists of 50 rounds. In each round, your earnings will be denoted in points. Your earnings in the experiment will be equal to the sum of the starting capital and your earnings in the 50 rounds. At the end of the experiment, your earnings in points will be transferred into money. For each 100 points you earn, you will receive 1 euro. Your earnings will be privately paid to you in cash.
In each of the 50 rounds of the experiment all participants are coupled in pairs. In each round you are randomly assigned to a new partner.
In each round, you and your partner will play a game in which you can choose between two options (**Cooperate** and **Defect**). Your earnings for this game will be determined by your and your partner's choice. After observing the choices in the game, the players choose whether or not to punish their partner. However, before the game each player decides how much punishment he or she can get at maximum. This number is referred to as 'the possible punishment level'. Both players are informed of the possible punishment levels in the own pair before they play the game.

INSTRUCTIONS PAGE 2

SEQUENCE OF EVENTS IN A ROUND
In the first phase of each round, you can determine how high punishment you might get later. You choose a possible punishment level that is at least 0 and at

most 50. In the experiment, a computerized clock raises the punishment level step-by-step starting from zero. If the clock shows the punishment level you would like to choose, you have to press the SUBMIT button, and this will be your possible punishment level later on in this round. At the same time your partner has the same procedure to decide about his or her possible punishment level. If either player presses the SUBMIT button, the other player is automatically and immediately informed about it. However, this phase does not end here, the other player has to press the SUBMIT button too. If a player does not press the SUBMIT button before the clock reaches the maximum level of 50, then the player's possible punishment level will be 50 in this round.
Notice that at the start of the second phase you and your partner are informed about each other's possible punishment levels. Then you will play a game where you and your partner have to choose between Cooperation and Defection. When you make your decision, you will not know your partner's decision (and neither does your partner know your decision). The earnings from this game will be determined as follows (where the first number in a cell determines your payoff, and the second your partner's payoff):

		Decision partner	
		C	D
Own decision	C	55,55	5,70
	D	70,5	25,25

In words, this means the following:
If both of you choose C, you and your partner will each earn 55 points.
If you choose C and your partner chooses D, you will earn 5 points and your partner earns 70 points.
If you choose D and your partner chooses C, you will earn 70 points and your partner earns 5 points.
If both of you choose D, you and your partner will each earn 25 points.
After both of you have decided, your choices will be revealed. In the third phase, you can decide whether you want to punish your partner. The punishment costs you 4 points. If you decide to punish, your partner gets a point deduction of the amount he or she determined him or herself in the first phase. If you decide not to punish, your partner does not get a point deduction, and you do not have to bear the punishment costs. At the same time, your partner also decides whether he or she wants to punish you. If he or she decides to punish you, he or she also bears a punishment cost of 4 points, and you get a point deduction of the amount you determined in the first phase. If he or she decides not to punish, you do not get a point deduction, and your partner does not have to bear the punishment costs.
After that the round is finished.

INSTRUCTIONS PAGE 3

ROUND EARNINGS
In each round, you can gain or lose points. Your earnings will consist of your payoff from the game minus the punishment you get from the other player minus the punishment costs if you punish your partner.
Note that your decision in the first phase (when you determine your punishment level) does not have a direct and immediate effect on your payoff. You get a point deduction only if your partner decides to punish you after observing your choice.

MATCHING PROCEDURE
In each round, you will be randomly matched to another participant. You will never learn with whom you are matched. The random matching scheme is chosen such that you will never be coupled to the same partner in two subsequent rounds.

INFORMATION
When you decide whether to cooperate or defect in the second phase, you know the possible punishment level your partner has chosen in the first phase. At the end of each round you will learn whether you received punishment or not. You will also be informed about the number of points you have earned in that round.

HISTORY OVERVIEW
The lower part of the screen provides an overview of the results of rounds already completed. If less than 10 rounds have been completed, this history overview contains results of all completed rounds. In case more than 10 rounds have already been completed, the history overview is restricted to the 10 most recent rounds.
The history overview contains the results of your group. The history overview contains your choices together with the choices of 7 other participants with whom you interact in the experiment. Sometimes your choices are listed as player 1's choices, sometimes as player 2's choices. Below you see an example of the history overview. The first column contains the possible punishment level of player 1 of a pair, whereas the second column shows the possible punishment level of player 2 of a pair. These punishment levels were determined in the first phase. The third column shows whether player 1 chose to cooperate or not. Column 4 shows whether player 2 chose to cooperate or not. Column 5 indicates whether player 1 received a punishment or not (yes: punishment, no: no punishment). Column 6 indicates whether player 2 received a punishment or not (yes: punishment, no: no punishment).

The past observations in the history screen have been ordered first by the punishment level of player 1. If these numbers are the same across pairs, these observations have been ordered by player 2's punishment level. If these numbers are also the same, then they are sorted on cooperation.
On the next screen you will be requested to answer some control questions. Please answer these questions now.

Player 1's punishment level	Player 2's punishment level	Player 1's decision	Player 2's decision	Was player 1 punished	Was player 2 punished
0	0	D	D	no	no
5	0	D	D	no	no
5	10	C	C	no	no
10	1	C	D	no	no
18	21	D	D	no	yes
20	10	C	D	no	yes
20	15	C	D	no	yes
25	30	C	C	no	no

References

Altman, I. (1973). Reciprocity of Interpersonal Exchange. *Journal for the Theory of Social Behaviour*, 3(2):249–261.

Andreoni, J. and Samuelson, L. (2006). Building rational cooperation. *Journal of Economic Theory*, 127(1):117–154.

Andreoni, J. and Varian, H. (1999). Preplay contracting in the Prisoners' Dilemma. *Proceedings of the National Academy of Sciences*, 96(19):10933–10938.

Ben-Porath, E. and Dekel, E. (1992). Signaling Future Actions and the Potential for Sacrifice. *Journal of Economic Theory*, 57(1):36–51.

Bolton, G. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Charness, G., Fréchette, G., and Qin, C.-Z. (2007). Endogenous transfers in the Prisoner's Dilemma game: An experimental test of cooperation and coordination. *Games and Economic Behavior*, 60(2):287–306.

Chen, Y. and Gazzale, R. (2004). When Does Learning in Games Generate Convergence to Nash Equilibria? The Role of Supermodularity in an Experimental Setting. *American Economic Review*, 94(5):1505–1535.

Coase, R. (1960). The Problem of Social Cost. *Journal of Law and Economics*, 3:1–44.

Derlega, V., Wilson, M., and Chaikin, A. (1976). Friendship and Disclosure Reciprocity. *Journal of Personality and Social Psychology*, 34(4):578–582.

Dindia, K. and Allen, M. (1992). Sex differences in self-disclosure: A meta-analysis. *Psychological Bulletin*, 112(1):106–124.

Falkinger, J. (1996). Efficient private provision of public goods by rewarding deviations from average. *Journal of Public Economics*, 62(3):413–422.

Falkinger, J., Fehr, E., Gächter, S., and Winter-Ebmer, R. (2000). A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence. *American Economic Review*, 90(1):247–264.

Fehr, E. and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fehr, E. and Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Gambetta, D. (2009). *Codes of the Underworld: How Criminals Communicate*. Princeton University Press.

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). Voting with Feet: Community Choice in Social Dilemmas. *IZA Discussion Paper*, No. 4643.

Hamaguchi, Y., Mitani, S., and Saijo, T. (2003). Does the Varian Mechanism Work? - Emissions Trading as an Example. *International Journal of Business and Economics*, 2(2):85–96.

Herrmann, I. and Palmieri, D. (2005). A haunting figure: The hostage through the ages. *International Review of the Red Cross*, 87(857):135–145.

Iossa, E. and Spagnolo, G. (2011). Contracts as Threats: on a Rationale For Rewarding A while Hoping For B. *CEPR Discussion Papers*, DP8195.

Kosfeld, M., Okada, A., and Riedl, A. (2009). Institution Formation in Public Goods Games. *American Economic Review*, 99(4):1335–1355.

Laurenceau, J.-P., Barrett, L., and Pietromonaco, P. (1998). Intimacy as an Interpersonal Process: The Importance of Self-Disclosure, Partner Disclosure, and Perceived Partner Responsiveness in Interpersonal Exchanges. *Journal of Personality and Social Psychology*, 74(5):1238–1251.

Lee, A. (1991). The Role of Hostages in Roman Diplomacy with Sasanian Persia. *Historia: Zeitschrift fr Alte Geschichte*, 40(3):366–374.

Levine, D. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3):593–622.

Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review*, 86(2):404–417.

Plutarch (1992). *Essays*. London: Penguin.

Potters, J. and Suetens, S. (2009). Cooperation in Experimental Games of Strategic Complements and Substitutes. *Review of Economic Studies*, 76(3):1125–1147.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83(5):1281–1302.

Raub, W. (2009). Commitments by Hostage Posting. In Baurmann, M. and Lahno, B., editors, *Perspectives in Moral Science*, pages 207–225. Frankfurt am Main: Frankfurt School Verlag.

Raub, W. and Keren, G. (1993). Hostages as a commitment device: A game-theoretic model and an empirical test of some scenarios. *Journal of Economic Behavior and Organization*, 21(1):43–67.

Rotenberg, K. (1986). Same-Sex Patterns and Sex Differences in The Trust-Value Basis of Children's Friendship. *Sex Roles*, 15(11):613–626.

Rtischev, D. (2011). Evolution of Vulnerability to Pain in Interpersonal Relations as a Strategic Trait Aiding Cooperation. *Journal of Evolutionary Economics*, 21(5):757–782.

Schelling, T. (1960). *The strategy of conflict*. Harvard University Press, Cambridge, Massachusetts.

Sefton, M., Shupp, R., and Walker, J. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45(4):671–690.

Snijders, C. and Buskens, V. (2001). How to convince someone that you can be trusted? The role of 'hostages'. *Journal of Mathematical Sociology*, 25(4):355–383.

Sommerfeld, R., Krambeck, H.-J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences*, 104(44):17435–17440.

Sutter, M., Haigner, S., and Kocher, M. (2010). Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies*, 77(4):1540–1566.

Varian, H. (1994). A Solution to the Problem of Externalities When Agents Are Well-Informed. *American Economic Review*, 84(5):1278–1293.

Weber, R. (2006). Managing Growth to Achieve Efficient Coordination in Large Groups. *American Economic Review*, 96(1):114–126.

Williamson, O. (1983). Credible Commitments: Using Hostages to Support Exchange. *American Economic Review*, 73(4):519–540.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51(1):110–116.