JENA ECONOMIC RESEARCH PAPERS

# 2013 – 016

Fairness norms can explain the emergence of specific cooperation norms in the Battle of the Prisoners Dilemma

by

Fabian Winter

www.jenecon.de

ISSN 1864-7057

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Fairness norms can explain the emergence of specific cooperation norms in the Battle of the Prisoners Dilemma

Fabian Winter[a]

[a] *Max Planck Institute of Economics, Kahlaische Straße 10, 07745 Jena, Germany, email: winter@econ.mpg.de*

Abstract

Cooperation norms often emerge in situations, where the long term collective benefits help to overcome short run individual interests, for instance in repeated Prisoner's Dilemma (PD) situations. Often, however, there are different paths to cooperation, benefiting different kinds of actors to different degrees. This leads to payoff asymmetries even in the state of cooperation, and consequently can give rise to normative conflicts about which norms should be in place. This norm-coordination problem will be modeled as a Battle of the Sexes game (BoS) with different degrees of asymmetry in payoffs. We combine the PD and the BoS to the 3×3 Battle of the Prisoners Dilemma (BOPD) with several asymmetric cooperative and one non-cooperative equilibria. Bame theoretical and "behavioral" predictions are derived about the kind of norms that are likely to emerge under different shadows of the future and degrees of asymmetry and tested in a lab-experiment. Our experimental data show that game theory fairly well predicts the basic main effects of our experimental manipulations, but "behavioral" predictions perform better in describing the equilibrium selection process of emerging norms.

Keywords: Social norms, normative conflict, Prisoner's Dilemma, coordination, experiment.
JEL-Classification: Z13, C92, C72, D31

1. Introduction

The problem of social order as well as the consequences of social inequalities have been a corner stone of sociological thought since the beginning of the discipline. While the diversity of actors and their interests' has widely been recognized, the lion share of attention in the Rational Choice literature has been gathered by explaining cooperation among symmetric actors on the one hand and social inequalities on the other. With the dominating focuss on symmetric actors, however, some important insights on the solutions in asymmetric games have widely been neglected (de Jasay et al., 2004).[1]

In this paper, we will shed light on the interaction between fairness norms, social inequality and the emergence of cooperation norms (see also Aksoy and Weesie, 2009). The novelty of our approach lies in the explicit investigation of the predictive power of fairness norms on the emergence of *different* cooperation norms. In this context, asymmetric games are particularly suited to model the existence of social inequality, which can lead to intriguing problems such as the emergence of normative conflict (Winter et al., forthcoming; Miller et al., 2011; Nikiforakis et al., 2011). This type of conflict exists not because people fail to overcome a cooperation problem, but because they can not agree on *which* norms should guide their behavior in order to overcome their collective obstacles. If we think of common fate problems, for instance a firm and it's workers, we can easily imagine this situation. Both, the owner of the firm and the worker, usually have an interest in a flourishing company. However, while the principal wants low wages and a high working level, the employees would prefer the opposite. Though norms are likely to emerge in these situations as well, the respective content of the norm has to be negotiated.

The experimental literature on asymmetric dilemmas has mainly focussed on the study of N-person public goods problems. In these games, every member of a group of players can decide to invest into a beneficial common project, which will then be distributed among them. Not investing, however, is usually a dominant strategy, which leads everybody worse-off as compared

[1] There are important differences in the kind of asymmetries that actors face such as the roles of leaders an followers or the access to information, to name just a few. Due to space restrictions, this paper will solely focuss on social inequalities as asymmetry in material well-being.

. *August 1, 2012*

to global investment.[2]. Different forms of social inequalities have been investigated in this context with mostly inconclusive results, such as different initial wealth levels leading to more resp. less cooperation (van Dijk and Wilke, 1995; Chan et al., 1996; Buckley and Croson, 2006; Kroll et al., 2007) or different marginal benefit from the public good leading to more resp. less cooperation (Glöckner et al., 2011; Reuben and Riedl, 2011). All these studies have in common that global efficiency is always at odds with a fair outcome: Full contribution to the public good necessarily leaves some actors better of than others. Moreover, the overwhelming majority of experimental studies investigates cooperation problems under a finite horizon, often accompanied by some form of a sanctioning mechanism (for seminal papers on off-equilibrium punishment see Yamagishi (1986); Ostrom et al. (1992) and for a study on equilibrium punishment see (Bruttel et al., 2009)).

This paper takes a different modeling approach to the cooperation problem. Instead of only one efficient cooperative solutions, there will be cases where several efficient cooperative equilibria are feasible, but only a subset of those are fair in the sense that they balance the monetary outcomes for the players. In what follows, we will shortly review how the cooperation problem can be overcome in indefinitely repeated games (section 2). Section 3 introduces a general game of asymmetric preferences and discusses how a cooperation problem could be solved if the actors coordinate on a shared norm. Section 4 derives hypotheses on the level of cooperation and on the evolving norms by means of game theory and fairness norms (section 5). Section 6 introduces our experimental design, section 7 presents the results and section 8 concludes.

2. Solving the cooperation problem in repeated interactions

The demand for social norms in repeated dilemma situation has extensively been studied in the theoretical Rational Choice literature (Ullmann-Margalit, 1977; Coleman, 1990). The effect of infinite or indefinite horizons on the emergence of cooperation norms, however, has largely been neglected by the experimentalists (see Gonzalez et al., 2005; Bruttel et al., 2007, for exeptions). This is surprising, as many (game theoretical) solutions to the cooperation problem rely on the fact that we face an infinite, or at least indefinite horizon (Taylor, 1976; Bicchieri, 1990; Ellickson, 1991; Voss, 2001). In fact, even many definitions of social norms rely on the fact that they emerge in repeated interactions:

Definition 1. ***(Social Norm** (Voss, 2001))*
A social [cooperation] *norm is a regularity* R *in a population* P *such that*

R *arises in recurrent interactions among the agents of population* P

almost every member of P *prefers to conform to* R *on the condition that almost every other member of* P *also conforms to* R

almost every member of P *believes that almost every other member of* P *conforms to* R

R *is a Nash equilibrium of the recurrent interaction.* [brackets added]

The theoretical possibility of cooperation gains emerges from the fact that norm breaking behavior can be reciprocated in future interactions (Gouldner, 1960). The Prisoners Dilemma (PD, top left of figure 1 on page 4) models such a symmetric cooperation problem among two players, both having a dominant strategy to defect which results in a socially undesired state of mutual defection. If the game is played repeatedly, the question here is how to surmount the players' myopic self-interest in order to achieve a mutually beneficial long-term cooperation. From a Rational Choice perspective, mere repetition of the PD does not solve the issue: Even in the finitely repeated PD, backward induction forces rational players into ruinous defection.

It is crucial, however, that the critical interactions are embedded in an indefinitely long repeated context. And in fact, many of our daily interactions are: We know that they will stop one day in the future, but we rarely know when this day will come exactly. This uncertainty about the *shadow of the future* gives rise to a whole class social norms which can pareto-improve the outcome of the social dilemma.[3]

[2] There are a few exceptions where at least one member has a dominant strategy to invest, see Marwell and Ames (1979) for a sociological and Reuben and Riedl (2011) for an economic contribution

[3] By formalizing the *folk theorem*, Fudenberg and Maskin (1986) can show that *any* sequence of actions can be supported by a Nash equilibrium in the infinitely repeated game without discounting. This seems rather farfetched in the context of social norms. It would translate into arbitrarily complex norms which lead to patterns of action that can last over a large number of interactions.

A long shadow of the future enables the actors to punish norm violations by doing as they are done by: Defection in one period can be repaid by defection in one or more subsequent periods. The most severe punishment strategies are triggered by a singular defection of one player, which leads to eternal defection by the other. We refer to this strategies as TRIGGER.

Definition 2. ***(TRIGGER)***
A TRIGGER strategy for repeated games reacts to a single defection with eternal defection.

It should be intuitively clear that TRIGGER poses an enormous threat on norm breaking behavior, where the size of the threat depends to a large extent on the likelihood δ that players attach to the event of playing another period.[4] The higher the actors evaluate the chances that they will meet again and the longer the time of punishment, the better the chances for cooperation. In this line of reasoning, TRIGGER sets the lower limit of the threat level. If the emergence of cooperation norms is not individually rational between actors using TRIGGER, it is not individually rational between actors using any other strategy (Abreu, 1988).

Lemma 1. ***(Equilibria in PD** (Axelrod, 1984))*
Mutual cooperation is a Nash equilibrium in the iterated PD for a pair of TRIGGER strategies and a discount parameter δ if

$$\delta > \frac{T-R}{T-P}$$

Proof: See appendix 9.

Note, that this result was shown for the classical PD, which is a *symmetric* game. Both players have aligned interests in switching from (D, D) to (C, C). As we have pointed out already in the introduction, this need not be the case in a more general framework.

3. Solving the "coordinate to cooperate" problem in repeated interaction

As we have already pointed earlier, cooperation problems often do not only face the problem of defection, but also the question of how to coordinate the distribution of the mutual benefits. We model this game as a generalization of the PD, flexible enough to capture symmetric as well as asymmetric solutions to the cooperation problem. Depending on the parametrization, This new game can have egalitarian equilibria as well as asymmetric solutions, leaving one player worse off. We will discuss the game shortly in general terms and introduce a parametrization which combines elements of the PD and the *Battle of the Sexes* game (see the top right side of figure 1. This new game will consequently be termed the Battle of the Prisoner's Dilemma (BOPD, see the bottom of figure 1).

The stage game of BOPD we will discuss here has the general payoff relations $T > a, b, c, d > P > S$ and $T > \alpha, \beta, \gamma, \theta > P > S$. The strategy combination (D, D) is the unique Nash equilibrium, but this equilibrium is pareto-dominated by some combination of C_1 and C_2, which constitutes the social dilemma. If we assume $a = b = c = d$ and $\alpha = \beta = \gamma = \theta$, the BOPD collapses to a PD with two instead of only one C-choice, such that the result from lemma 1 still holds.

3.1. The coordination problem

There are several ways how two players could cooperate in these kind of games, but only few of them are theoretically feasible and still sufficiently simple to assume that they could reasonably emerge. We therefore concentrate on the two behavioral norms of cooperation in pure strategies and the norm of turn taking.

Let us shortly focus on the C_1 and C_2 options of BOPD, which constitute a Battle of the Sexes game if we assume that $\alpha > \theta > \beta, \gamma$ and $d > a > b, c$. Both players have an incentive to coordinate, either on C_1, C_1 or on C_2, C_2. However, while the row player prefers C_1, C_1, the collum player would like to coordinate on C_2, C_2. Moreover, either of the two possibilities would

[4] In this paper, we use the two concepts of uncertainty about future interactions and the discounting of future gains as compared to the present ones interchangeably. Usually, discounting is applied to solve *infinite* games, whereas a fixed termination probability makes a game *indefinite*. Note, that discounting can also take place in indefinite games (Vogt and Weesie, 2004).

Prisoner's Dilemma

Col Row	C	D
C	R R	T S
D	S T	P P

$T > R > P > S$

Battle of the Sexes

Col Row	C_1	C_2
C_1	$\alpha = 1$ a = 2	$\beta = 0$ b=0
C_2	$\gamma = 0$ c=0	$\theta = 2$ d=1

$d > a > b, c$ and $\alpha > \theta > \beta, \gamma$

Battle of the Prisoners Dilemma

Col Row	C_1	C_2	D
C_1	α a	β b	T S
C_2	γ c	θ d	T S
D	S T	S T	P P

$T > d > a > b, c > P > S$
$T > \alpha > \theta > \beta, \gamma > P > S$

BOPD, numeric example

Col Row	C_1	C_2	D
C_1	50 80	40 40	100 0
C_2	40 40	80 50	100 0
D	0 100	0 100	30 30

Figure 1. The Prisoner's Dilemma (PD, top left), the Battle of the Sexes (BoS, bottom left), and the Battle of the Prisoner's Dilemma (BOPD, right). In BOPD, the C option of the PD is substituted by two new options C_1 and C_2 containing a BoS.

make one player worse of than the other. Obviously, coordination is a Nash equilibrium in pure strategies, yielding the pay-off

$$\pi_{pure} = \begin{cases} \min(a, \theta), & \text{to the worse-off player, and} \\ \max(d, \alpha), & \text{to the better-off player.} \end{cases} \tag{1}$$

There exists another equilibrium in correlated strategies. In the absence of an exogenous signaling mechanism, this equilibrium is not feasible in the the stage game. In repeated games, however, subjects can endogenously "signal to each other via their choice patterns on previous plays. Introspectively, we would suspect that, after some preliminary jockeying, the players would settle on a pattern of alternation"(Luce and Raiffa, 1989, 94). If both players manage to coordinate on jointly alternating between C_1 and C_2, this *turn taking* yields a expected payoff of

$$\pi_c = \begin{cases} \dfrac{a+d}{2}, & \text{to the row player, and} \\ \dfrac{\alpha+\theta}{2}, & \text{to the collumn player.} \end{cases} \tag{2}$$

The correlated equilibrium's outcome is always between the two outcomes in pure strategies.[5] Note, that all cooperative equilibria in BOPD are payoff asymmetric in the one-shot game. The cooperative pure strategies equilibria, however, are payoff asymmetric even in the repeated game, while the correlated strategies can be symmetric, depending on the payoff matrix.

[5] There exists an additional equilibrium in mixed strategies, but we will refrain from the discussion of the mixed strategy as a candidate for a social norm implausibility reasons. The mixed strategy equilibrium requires that players to mix their strategies such that the other player is indifferent between playing C_1 and C_2. Players calculate an optimal mixing proportion between the two actions as a function of the other player's payoff:

3.2. Feasible norms in the Battle of the Prisoners Dilemma

Under which conditions can these different equilibria emerge as norms in the Battle of the Prisoner's Dilemma? Remember that TRIGGER sets the shadow of the future's lower bound by immediately reacting on norm breaking behavior. We could, however, think of more complex TRIGGER-strategies, which do not simply react on a deviation from C, but rather on the deviation from a *cooperative pattern*, such as turn taking.

Definition 3. ***(TRIGGER*)***
A TRIGGER strategy reacts on any deviation from a cooperative pattern with eternal defection, where a cooperative pattern is some combination* ε *of moves in at least one instance of a super-game.*

TRIGGER*-strategies set the shadow of the future's lower bound for cooperative strategies other than pure cooperation. We can apply the same analysis as for the PD and calculate the lower bound by taking the expected long term outcome of the cooperative pattern as the benchmark:

Lemma 2. *(Equilibria in BOPD)*
Let $\pi^*(\varepsilon)$ *be player i's expected pay-off from some combination* ε *of moves in at least one instance of a super-game of BOPD games. Then* ε *is a Nash equilibrium in the iterated BOPD for a pair of TRIGGER* strategies if*

$$\forall i: \quad \delta > \frac{T-\pi^*}{T-P}.$$

Proof: See appendix 9.

The possibility of cooperation depends exclusively on the discount factor and the worse off player's temptation to defect. Even if the better-off player would prefer to cooperate at a given discount factor, this cannot be an equilibrium if the other player has an incentive to defect. Equation 4 reports the respective worse-off player's pay-offs for different cooperation norms in BOPD:

$$\pi^* = \begin{cases} \min(\pi_p), & \text{pure strategies} \\ \min(\pi_c), & \text{turn taking} \end{cases} \tag{4}$$

If the repeated gains from mutual cooperation can outweigh the temptation of a singular exploitation, it can be rational to stick to the cooperative solution than to risk the unpleasant impact of defections.

Note, that lemma 2 can also be applied to the PD. If we assume $(T+S)/2 > R$, turn taking yields a higher long-run pay-off in PD as compared cooperation in pure C-strategies. The critical discount factor δ for this cooperative norm is thus given by

$$\delta > \frac{T-\frac{T+S}{2}}{T-P} > \frac{T-R}{T-P}, \tag{5}$$

which is a shorter shadow of the future than the one necessary for cooperation in pure strategies.

4. Hypothesis on the emergence of cooperation norms

We will now propose a parametrization for the PD and the BOPD in order to derive hypothesis which we will later test in an experiment. Figure 1 on 4 shows four different games, the symmetric and asymmetric PD (top row) and the symmetric and asymmetric BOPD (bottom row). The pay-off symmetric game can be transformed to the pay-off asymmetric game by adding 10 to the row-player's pay-offs in the C_i-cells. In the PD, the pay-off from pure C-strategies is 40 (or 50 for the better-off player in the asymmetric PD), which is below or equal to the expected gains from turn taking between C and D, returning $(100+0)/2 = 50$. This gives rise to an efficient equilibrium in turn taking strategies also in the symmetric as well as the asymmetric PD.

$$p = \begin{cases} \dfrac{d-b}{a-b+d-c}, & \text{for the row player, and} \\ \dfrac{\theta-\gamma}{\alpha-\beta+\theta-\gamma}, & \text{for the collumn player,} \end{cases} \tag{3}$$

The expected payoff from the mixed equilibrium is always below the payoff obtained from pure strategies, even if a player happens to agree on the worse option, but it is a *save option*, as it gives the same expected payoff independent of the other player's action.

Consider now the BOPD in bottom row of figure 2. Here, coordinating on the C_1, C_1 yields a pay-off of 80(50) to the row(collum)-player, while C_2, C_2 yields the reversed payoff of 50(80). Thus, coordinating on a pure C_1, C_1 or C_2, C_2 equilibrium is efficiency enhancing but lets one player always worse-off as compared to other. In contrast to that, jointly alternating between both cells will equalize the payoffs after every even period in the symmetric BOPD, returning (50+80)/2. Moreover, turn taking can minimize the pay-off differences even in the asymmetric BOPD as it yields (60+90)/2=75 to the row-player and (50+80)/2=65 to the column-player, which is less than the minimal difference in pure strategies of 20. Note, that coordination on one cell as well as jointly alternating is socially efficient, such that the only question here is how to distribute the surplus.

The table in figure 2 reports critical discount factors $\underline{\delta}$ for the emergence of different cooperation norm. These discount factors are regarded necessary but not sufficient conditions for the emergence of cooperation. Moreover, they can also tell us which norms are likely to emerge under different pay-off structures. Asymmetric pay-offs, for instance, have a positive effect on the emergence of a pure (C_2, C_2) norm in the BOPD, as the discount factor falls from .71 to .5. The same reasoning leads us to predict a positive effect of the shadow on the future on the emergence of pure strategy norms. We can now derive hypothesis about the ways people cooperate in symmetric and asymmetric social dilemmas.

Hypothesis 1. ***(Games)***
There is more

a *cooperation in the* BOPD *as compared to the* PD*.*

b *turn taking in the* BOPD *as compared to the* PD*.*

c *pure strategy behavior in the* BOPD *as compared to the* PD*.*

Hypothesis 2. ***(Shadow of the future)***
A long shadow of the future has

a *a* **positive effect** *on the emergence of cooperation in the* BOPD *and a* **positive effect** *in* PD*.*

b **no effect** *on the emergence of turn taking in the* BOPD *and* **no effect** *in the* PD*.*

c *a* **positive effect** *on the emergence of pure strategy norms in* BOPD *and* **positive effect** *in* PD*.*

Hypothesis 3. ***(Asymmetry)***
Asymmetric pay-offs have

a *a* **positive effect** *on the emergence of cooperation in the* BOPD *and* **no effect** *in* PD*.*

b **no effect** *on the emergence of turn taking in the* BOPD *and* **no effect** *in the* PD*.*

c *a* **positive effect** *on the emergence of pure strategy norms in* BOPD *and* **no effect** *in* PD*.*

5. Fairness norms as a predictor for the emergence of different cooperation norms.

In addition to the cooperation problem, the players in the games investigated here face an allocation problem. Evidently, cooperation can lead to efficiency gains in the BOPD as well as in the PD. As the pay-offs are asymmetric, however, the fruits of cooperation can be distributed in different ways. How can fairness-norms deepen our understanding of the emergence of specific cooperation norms?

An important contribution to solve these question has been put forth by a stream of literature on “social preferences” or fairness norms (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In this line of thought, individual utility does not only arise from material outcomes, but additionally from the comparison of own and other peoples pay-offs. Depending on the respective value subjects assign to each of these parameters, some subjects exclusively maximize their own payoffs (individualistic maximizers), minimize the distance between the own and the other player‘s pay-off (prosocial), maximize the distance between the own and the other player‘s

symmetric **asymmetric**

2 × 2 PD

Col Row	C	D
C	40 40	100 0
D	0 100	30 30

Col Row	C	D
C	40 50	100 0
D	0 100	30 30

3 × 3 BOPD

Col Row	C_1	C_2	D
C_1	50 80	40 40	100 0
C_2	40 40	80 50	100 0
D	0 100	0 100	30 30

Col Row	C_1	C_2	D
C_1	50 90	40 50	100 0
C_2	40 50	80 60	100 0
D	0 100	0 100	30 30

symmetric 2×2 PD **asymmetric 2×2 PD**

	necessary	equilibrium for		necessary	equilibrium for	
strategy	$\underline{\delta}$	$\delta = .7$	$\delta = .9$	$\underline{\delta}$	$\delta = .7$	$\delta = .9$
pure (C, C)	.85	−	+	.85	−	+
turn taking	.71	+	+	.71	+	+

symmetric 3×3 BOPD **asymmetric 3×3 BOPD**

	necessary	equilibrium for		necessary	equilibrium for	
strategy	$\underline{\delta}$	$\delta = .7$	$\delta = .9$	$\underline{\delta}$	$\delta = .7$	$\delta = .9$
pure (C_1, C_1)	.71	−	+	.71	−	+
pure (C_2, C_2)	.71	−	+	.57	+	+
turn taking	.5	+	+	.5	+	+

Figure 2. Four different social dilemmas investigated in the experiment. The symmetric Prisoner's Dilemma (top left), the asymmetric Prisoner's Dilemma (top right), the symmetric 3×3 Battle of the Prisoner's Dilemma (bottom left) and the asymmetric 3×3 Battle of the Prisoner's Dilemma. All games are dilemmas, as mutual defection (D, D) is a dominant strategy. The BOPD extends the PD by a Battle of the Sexes game in the cells C_1 and C_2 if actors want to move from mutual defection towards some form of cooperation. The symmetric game can be transformed to the asymmetric game by adding 10 to the row-player's pay-offs in the C-cells. The table on the bottom displays the necessary discount factors $\underline{\delta}$ for different strategies to be an equilibrium in all four games.

Treatment	Game	Symmetric/Asymmetric	Discount Factor	N
2×2SymLow	PD	symmetric	.7	30
2×2SymHigh	PD	symmetric	.9	30
2×2AsymLow	PD	asymmetric	.7	30
2×2AsymHigh	PD	asymmetric	.9	30
3×3SymLow	BOPD	symmetric	.7	30
3×3SymHigh	BOPD	symmetric	.9	30
3×3AsymLow	BOPD	asymmetric	.7	30
3×3AsymHigh	BOPD	asymmetric	.9	30

Table 1. Treatment Conditions

pay-off (competitive), or the other player's outcome (altruists). As maximizers and prosocials are the most common types (Murphy et al., 2011), we focus exclusively on these two.[6]

Hypothesis 4. ***(Fairness norms)***
As compared to individualistic maximizers, the emergence of

a *cooperation is* **positively affected** *by fairness norms.*

b *turn taking norms is* **positively affected** *by fairness norms.*

c *pure strategy norms is* **not affected** *by fairness norms.*

Although cooperation helps both types to maximize their earnings, individualistic types are probably more tempted to play the dominant strategy of defection, as they would not want to risk to leave with the sucker's pay-off. Hence, subjects adhering to fairness norms can be expected to cooperate more than individualistic types. Similarly, as prosocials derive utility from balanced pay-offs, turn taking can be expected to be more prominent among them. Finally, pure strategy norms have different effects for different games and asymmetries on prosocial types, also depending on their commitment to the norm. They would for instance prefer cooperation over defection in the symmetric PD, but for sufficiently large fairness concerns defection over cooperation in the asymmetric case. We consequently predict similar probabilities for the emergence of pure strategy norms between individualistic and prosocial types.

6. Methods

We designed a lab experiment in order to test the predictions derived for the four games discussed in table 2. The experiment was conducted in an incentive compatible manner using the *z-Tree* software developed by Fischbacher (2007). Our experimental subjects were 240 undergraduate students from a large European university, recruited from a wide range of academic disciplines with the online recruiting system ORSEE (Greiner, 2004). 93 subjects were male and 247 female.

6.1. Experimental design

We employed a $2 \times 2 \times 2$ factorial between subjects design in which we manipulated the shadow of the future ($\delta = .7$ and $\delta = .9$), the asymmetry between payoffs for the players (pay-off symmetric and asymmetric players) and the coordination problem (2×2 PD and 3×3 BOPD, for an overview see figure 2 and table 1).

Before the experiment, the instructions were presented on the computer screen[7], and were intended to familiarize the subjects with the game matrix and the concept of the shadow of the

[6] A frequently discussed model of such a utility function was proposed by Fehr and Schmidt (1999), which has the following form:

$$U_i(x) = x_i - \alpha_i \max\left[x_j - x_i, 0\right] - \beta_i \max\left[x_i - x_j, 0\right],$$

where x_i is the player's pay-off, x_j is the other player's pay-off, α_i is a parameter of "envy" and β_i a parameter of "guilt". The parameters α_i and β_i are usually assumed to be positive, which restricts the model to prosocials and maximizers. Altruists could be modeled by assuming negative values for α, as they actually derive utility from the other person being better-off. Conversely, competitive types derive utility from other people being worse-off, which could be modeled by a negative β_i. See also Tutic and Liebe (2009) for a related modeling approach and Aksoy and Weesie (2009) for a model and experimental evidence.

[7] We used the software E-nstructions, see Schmelz (2010)

endpoints 85, 85 and 85, 15). A perfectly consistent competitor yields an angle of -16.26°. Given

8: This figure shows where in the self/other allocation plane the six primary items are from the Measure.

Figure 9: This figure shows the location of the nine secondary items for the Slider Measure in t self/other allocation plane.

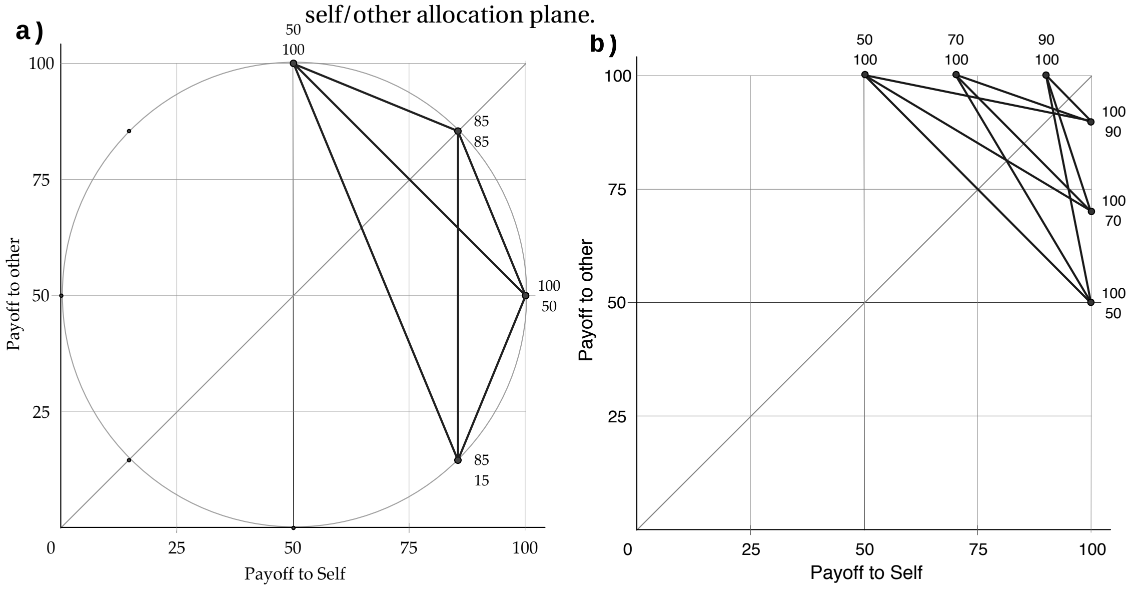

Figure 3. The social value orientation slider measure (figure taken from Murphy et al. (2011)). The six primary items in a) elicit general social value orientations, while the nine secondary items in b) allow to discriminate between concerns for equality and efficiency. The bold lines represent the decision tasks with bold circles at the end-points, for instance the decision between any point on the line 50/100 for player i/j and 85/85 for i/j (top right decision in a)), or the decision between 85/85 for i/j and 85/15 for i/j (the vertical bold line on the right in a)). The bold circles of the primary items represent the 4 ideal-typical social value orientations: altruistic (50/100), prosocial (85/85), individualistic (100/50), and competitive (85/15).

future (figure 6 in appendix 9 was presented to the subjects to illustrate the prospect of future interactions). Subjects had to answer some control questions in order to assure that everybody understood the setup of the experiment.

6.2. Phase 1: Social dilemma game

In phase 1 of the experiment, every subject played four super-games with constant discount factor, asymmetry and coordination problem. After every super-game the subjects were rematched to another parter they had not played with before, and who had not played with anyone the subject had played with before.[10] A super-game consisted of several repetitions of the same game until a random draw terminated the game with a probability of $1-\delta$. The subjects were informed in the instructions that a game could alternatively be terminated if it takes "too long", which luckily was never necessary. The subjects had to choose between three resp. two different options presented as a normal form game. We presented the games such that the experimental subject always had to make the decisions as the row player and that the cooperative option with the highest pay-off was in the upper left cell. Subjects received feedback about the decisions of both players as well as a history of previous decisions in the same super-game.

6.3. Phase 2: Social value orientation

In phase 2 of the experiment, we elicited social value orientation using the social value orientation slider measure (see Murphy et al. (2011) for a very detailed discussion and (Rauhut and Winter, 2010) for a different approach on measuring fairness norms). This device is a measure to distinguish between very different kinds of motivations, such as altruistic, prosocial, individualistic, or competitive norms. It can be regarded a derivative of the ring measure introduced by Liebrand and McClintock (1988), yet it is less prone to inconsistent choices. The subjects were asked to play a series of dictator games between themselves and some other person in the room they have not interacted with before, which allows us to classify them into types. Figure 3 gives an overview about allocation decisions the subjects where facing. After they had submitted their decisions, one of the allocations of one of the two interacting players was chosen for payment. The results of this task will be presented elsewhere.

In order to prevent hedging across games, subjects where payed only one super-game from phase 1 plus the outcome of phase 2.

gles that result from idealized SVO types, proper boundaries between SVO categories can be determined (Slider primary items and/or Slider secondary items); second t

s would have an angle greater than 57.15°; prosocials should have angles between 22.45° and ... These subjects then receive an email invitation to pa

individualists would have scores between -12.04° and 22.45°; and competitive types would ... Each email invitation has a unique tokened link that when clicked upo

n angle less than -12.04°. As it can be seen, these boundaries are not at intuitive locations. The takes the subject directly to the SVO website. Once there, the subject is presented with instructio

for this is that the Slider Measure only uses a subset of possible items from the allocation plane and then a series of allocation choices. For each subject, the items are presented in a random ord

ese items are not symmetrically distributed around the whole of the ring. Because only the Subjects move the slider back and forth, changing the joint allocations un

they find their most preferred outcome (see Figure 10 for a screen shot). Once they are satisfied wi

the distribution, they finalize it by clicking the "Submit" button. Then the program takes them

the next item. If they have answered all of the items, to a conclusion and "thanks for participatin

page. After all the subjects on the invitation list have participated (or a predetermined end date/tim

whichever happens first) the researcher is sent via email a datafile with the subjects' identifying info

mation, item order, all of their choices, summary scores, and transitivity verification.

7. Results

7.1. Patterns of cooperation

We start our analysis with a qualitative look on the data in figure 4. The left column depicts interactions in the 2×2 Prisoner's Dilemma, the right column interactions from the 3×3 Battle of the Prisoner's Dilemma. In the top row the subjects coordinated on turn-taking, either between C, D and D, C in the normal PD, or between C_1, C_1 and C_2, C_2 in the BOPD. In fact, if subjects manage to cooperate over a long time horizon, turn taking is relatively common in the 2×2 PD game as we will see later. The reason for that may be that alternating yields an average pay-off of 50 ECU per period, whereas pure C, C play yields only 40 ECU (or 50 ECU to the row-player in the asymmetric case). The top-right panel gives a paradigmatic impression of a short normative conflict which is eventually resolved after period three: The column-player tries to establish a pure C_2-norm (which would be better for him than turn-taking or C_1), while the row-player starts with turn-taking. Finally, the column-player seems to agree on "turn-taking" and follows the row-player until the interaction ends.

The central panel is an example of coordination on pure strategies of cooperation, that is C, C in the PD and C_2, C_2 in the BOPD. Note that the interaction in BOPD took place in the asymmetric version of the game. The row player constantly plays C_2, which is the less preferred cooperative solution for him (60 ECU instead of 90 ECU), but more than the 50 ECU the column-player could expect from playing C_1. Thus, C_2 could be interpreted as a "friendly offer", which makes the cooperation stable over time.

The bottom panel is an example of how initially cooperative intentions are undermined by a unilateral choice of D. The row-player's behavior in the PD is consistent with a *TRIGGER*-strategy: He pays the column-player's first-round defection with defection in all subsequent rounds, notwithstanding column's tentative initiatives to re-establish cooperation. The bottom-right interaction in the BOPD is an other interesting instance of normative conflict: The row-player tries to force his partner into playing C_1, who himself rather would like to establish the turn-taking norm, which he seemingly tries to enforce by a one-time defection. This defection is in turn reacted upon by a D-choice of the row-player. However, after some more back and forth all efforts to establish a cooperation norm fail and both players choose D in almost all future periods.

7.2. Which cooperation norms emerge in the PD and in the BOPD?

We continue by investigating the different forms of cooperation by means of a series of random effects logistic regressions. The general model to be estimated is given by

$$\text{logit}\{Pr(y_{ij} = 1|\mathbf{x}_{ij})\} = \frac{\exp(\mathbf{x}_i\beta)}{\mathbf{1} + \exp(\mathbf{x}_i\beta)}, \tag{6}$$

where

$$\mathbf{x}_{ij}\beta = \beta_\mathbf{0} + \beta_\mathbf{1}\text{game} + \beta_\mathbf{2}\text{delta} + \beta_\mathbf{3}\text{asymmetric} + \beta_\mathbf{4}\text{period} + \nu_\mathbf{j}, \tag{7}$$

and β_0 is the intercept, game is a dummy taking the value 1 if subjects play the 3 × 3 BOPD game, delta is a dummy indicating that the shadow of the future is long, period is the respective interaction the subject is in, and asymmetric is a dummy taking the value 1 if the game to be played is asymmetric. The subject specific random intercept is denoted by ν_j and assumed to be drawn from the distribution $\text{N}(0, \psi)$. This random intercept accounts for the fact that the choices of one person are likely to be correlated, which would violate the assumption of uncorrelated errors in standard logistic regressions (Snijders and Bosker, 1999; Rabe-Hesketh and Skrondal, 2005). Depending on the analysis, the dependent variable y_ij is given by unilateral cooperation, or bilateral forms of cooperation, such as the joint probability of playing the same pure strategy or alternating between two choices. All estimates in the following analysis rely on bootstrapped parameters and confidence intervals, the reason being that confidence intervals have been shown to be more robust than p-values and bootstrapped confidence intervals have been shown to be more robust than those analytically derived (Efron, 1987).[8]

[8] The method of bootstrapping draws B sub-samples of size N with replacement from the data and estimates the respective model for every sub-sample (where B is in our case 1000 and N the number of observations in the data). The reported coefficients are the arithmetic means of the B bootstrapped coefficients of every independent variable, while the confidence intervals are the observed "inner" 95 % around the mean of the respective coefficients. See Efron and Tibshirani (1993) for an introduction to the bootstrap.

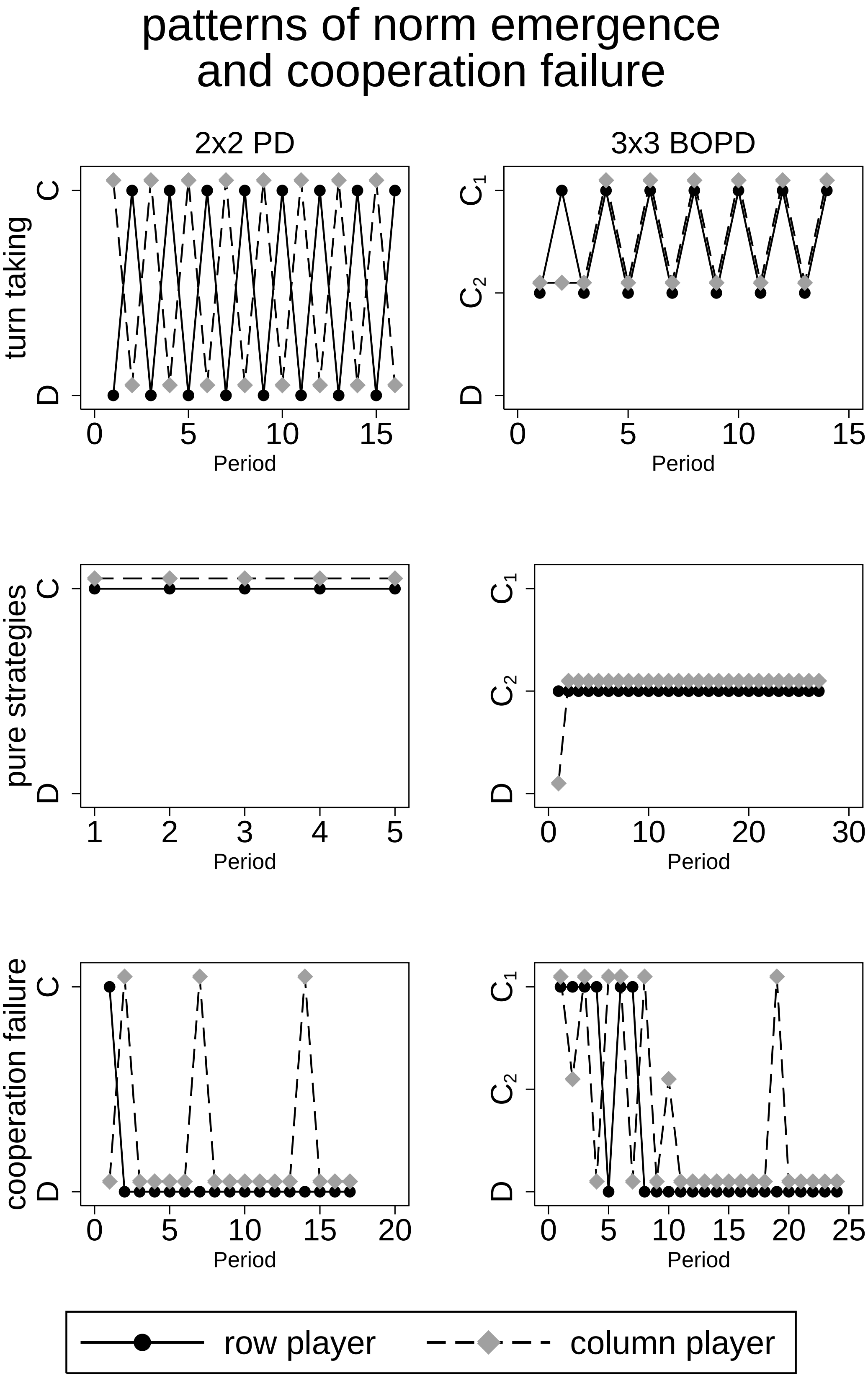

Figure 4. Some representative interactions in the experiment. The left column depicts interactions in the BOPD, the right column interactions from the PD. In the top row the subjects coordinated on turn-taking between C_1, C_1 and C_2, C_2 (left) and C, D and D, C (right). The middle row is an example of coordination on a pure strategy of cooperation, C_2, C_2 in BOPD and C, C in the PD game. The bottom row shows how initially cooperative intentions are undermined by a unilateral choice of D.

		cooperation		cooperation		cooperation
	sign	general	sign	turn taking	sign	pure strategies
fixed effects						
game (3×3=1)	+	1.178***	+	-2.986***	+	0.949***
		[0.995,1.360]		[-3.342,-2.629]		[0.470,1.428]
high discount factor	+	1.011***	0	1.898***	+	1.610***
		[0.813,1.209]		[1.439,2.357]		[0.783,2.436]
asymetric	+	0.0873	0	0.508**	+	0.266
		[-0.0926,0.267]		[0.172,0.844]		[-0.223,0.754]
period	−	-0.0621***	−	0.0435***	−	0.0138
		[-0.0757,-0.0484]		[0.0212,0.0658]		[-0.0102,0.0379]
intercept		-2.690***		-5.826***		-8.418***
		[-2.919,-2.461]		[-6.403,-5.249]		[-9.583,-7.252]
random effects						
var(intercept)		0.969***		1.923***		2.041***
		[0.825,1.113]		[1.759,2.087]		[1.824,2.258]
decisions		6078		6078		6078
subjects		240		240		240

bootstrapped 95% confidence intervals in brackets,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. Different forms of cooperation are likely to emerge in different games. All models are random effects logistic regressions controlling for correlated errors on the subject level (see equation 7). The first model on the left predicts the general degree of cooperation (i.e. non-D choices). The second model investigates the joint probabilities of turn taking between two choices, whereas the model on the right predicts cooperation in pure strategies. The "sign"-columns list the theoretically expected signs. Coefficients and confidence intervals are based on 1000 bootstraps.

Result 1. ***(Games)***
There is

a **more** *cooperation in the* BOPD *as compared to the* PD*.*

b **less** *turn taking in the* BOPD *as compared to the* PD*.*

c **more** *pure strategy behavior in the* BOPD *as compared to the* PD*.*

We start testing our hypotheses with a look at the overall level of cooperative choices (see the left model in table 2. We define a cooperative choice as an off-equilibrium choice in the finite game, that is playing C in the PD and C_1 or C_2 in the BOPD. The respective dummy variable takes the value 0 if the subject chooses D and 1 otherwise. We can confirm our hypothesis 1a concerning the higher level of cooperation in the BOPD as compared to the PD by a positive coefficient for the variable "game" in the left model of table 2. In fact, there may be several reasons for this finding. If we assume some subjects to behave purely random, the cooperation rate should still be higher in the BOPD, as two instead of only one choice is classified as cooperative. More reasonable, however, is the fact that subjects react to the lower opportunity costs of cooperation: While unilateral defection returns 100 ECU, cooperation in the PD returns 40 resp. 50 ECU, but up to 90 ECU in the BOPD.

In order to test how our experimental manipulations affect the emergence of one or the other norm, we generated two dummy variables, "turn taking" and "pure strategies", and use them as dependent variables in equation 7. "Turn taking" in the BOPD in some period t takes the value 1 if and only if both players' decision in period t is either C_1 or C_2, and C_2 or C_1 respectively in the preceding period t_{-1}. For the PD, it takes the value 1 in period t if and only if the first player chooses C and the second player chooses D in period t, whereas the first player had chosen D and the second player had chosen C in t_{-1}. A cooperation norm in pure strategies is present if both players made the same cooperative choice for two consecutive periods. More formally, "pure strategies" is a dummy variable, taking the value 1 in period t if and only if both players chose C in t and t_{-1} in PD, or C_1 in t and t_{-1} or C_2 in t and t_{-1} in BOPD.

Hypothesis 1b and c predicted more turn taking *and* more pure strategy play in the BOPD as compared to the PD. These hypotheses, however, can only partially be confirmed: The positive estimated coefficient for "game" in the second model of table 2 confirms the theoretically predicted higher propensity of pure strategy play in the BOPD. However, the Rational Choice theory fails to predict the significantly lower frequency of turn taking in the BOPD as compared to the PD, as we can see from the negative estimate for "game" in the third model.

7.3. Which cooperation norms emerge under the shadow of the future?

Our second experimental manipulation varied the shadow of the future.

Result 2. ***(Shadow of the Future)***
A long shadow of the future has

a *a* **positive effect** *on the emergence of cooperation in the* BOPD *and a* **positive effect** *in* PD*.*

b *a* **positive effect** *on the emergence of turn taking in the* BOPD *and* **no effect** *in the* PD*.*

c *a* **negative effect** *on the emergence of pure strategy norms in* BOPD *and a* **positive effect** *in* PD*.*

Hypothesis 2a is supported by a positive coefficient for the dummy "high discount factor" in the first model of table 2. Note, that the model controls for the fact that interactions under a short shadow of the future are on average shorter by adding the respective period to the regression. Moreover, this result is robust for both games, as can be concluded from the first two decomposed models in table 3.

Again, our the hypotheses concerning the emergence of turn taking or pure strategy are mostly inconsistent with the data. We predicted a nil-effect for the emergence of turn taking, but pooling the data from both games returns a significantly positive estimate for the shadow of the future (second model in table 2). The effect is mainly driven by the PD game, where turn taking norms are more likely if the shadow of the future is long. As predicted, we cannot find evidence for the same effect in the BOPD game (see models 3 and 4 in table 3).

7.4. Which cooperation norms emerge under asymmetric pay-offs?

Finally, asymmetric pay-offs do not play a role for the emergence of cooperation which can be referred from insignificant estimate for asymmetry (see first model in table 2). Decomposing both games, however, shows that this result is inconsistent across games (see the first two models in table 3). In fact, asymmetry has a negative effect in the PD, but a positive effect in the BOPD.

Result 3. ***(Asymmetry)***
Asymmetric pay-offs have

a *a* **positive effect** *on the emergence of cooperation in the* BOPD *and a* **negative effect** *in* PD*.*

b **positive effect** *on the emergence of turn taking in the* BOPD *and* **no effect** *in the* PD*.*

c *a* **negative effect** *on the emergence of pure strategy norms in* BOPD *and* **positive effect** *in* PD*.*

How do asymmetric pay-offs affect the emergence of turn taking norms? In contrast to our prediction, the regression for the pooled data estimates a positive relationship between asymmetry and the emergence of turn taking. The third and fourth model in table 3 estimate the effect of the experimental manipulations separately for both games. The Rational Choice theory predicts no difference between symmetric and asymmetric games (see hypothesis b), however, this effect is estimated to be positive for the BOPD, while it is not different from zero in the PD.

We can observe an almost reversed pattern for the emergence of pure-strategy norms: In contradiction to hypothesis 3c, asymmetric incentives promote the emergence of pure strategy play in the PD, but hinder their evolution in the more complex BOPD, such that the pooled effect is not different from zero (see models 5 and 6 in table 3 and model 3 in table 2).

	cooperation				turn taking				pure strategies			
	sign	2×2 PD	sign	3×3 BOPD	sign	2×2 PD	sign	3×3 BOPD	sign	2×2 PD	sign	3×3 BOPD
high δ	+	0.52***	+	1.44***	0	1.81***	0	2.33	+	21.30***	+	-0.99*
		[0.21,0.82]		[1.17,1.71]		[1.31,2.32]		[-2.77,7.42]		[10.77,31.82]		[-1.97,-0.00]
asymetric	0	-0.49***	+	0.59***	0	0.13	0	1.50***	0	1.85***	+	-1.76**
		[-0.78,-0.21]		[0.36,0.82]		[-0.28,0.54]		[1.04,1.97]		[1.39,2.31]		[-2.87,-0.65]
period	−	-0.06***	−	-0.06***	−	0.04**	−	0.05**	−	0.02	−	-0.01
		[-0.08,-0.03]		[-0.08,-0.05]		[0.01,0.07]		[0.02,0.08]		[-0.01,0.04]		[-0.09,0.08]
intercept		-2.22***		-1.95***		-5.47***		-10.25***		-27.11***		-5.97***
		[-2.53,-1.91]		[-2.23,-1.67]		[-6.12,-4.81]		[-15.32,-5.19]		[-37.63,-16.59]		[-7.61,-4.33]
random effects												
var(intercept)		1.14***		0.72***		1.81***		2.21***		1.62***		2.03***
		[0.92,1.36]		[0.53,0.91]		[1.61,2.01]		[1.93,2.48]		[1.34,1.89]		[1.55,2.52]
decisions		2862		3216		2862		3216		3216		2862
subjects		120		120		120		120		120		120

bootstrapped 95% confidence intervals in brackets,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Emergence of cooperation norms decomposed by games (PD and BOPD). Different forms of cooperation are likely to emerge in different games. All models are random effects logistic regressions controlling for correlated errors on the subject level (see equation 7). The first two models on the left predict the general degree of cooperation (i.e. non-D choices). The second block of models investigates the joint probabilities of turn taking between two choices, whereas the two models on the right predict cooperation in pure strategies. The "sign"-columns list the theoretically expected signs. Coefficients and confidence intervals are based on 1000 bootstraps.

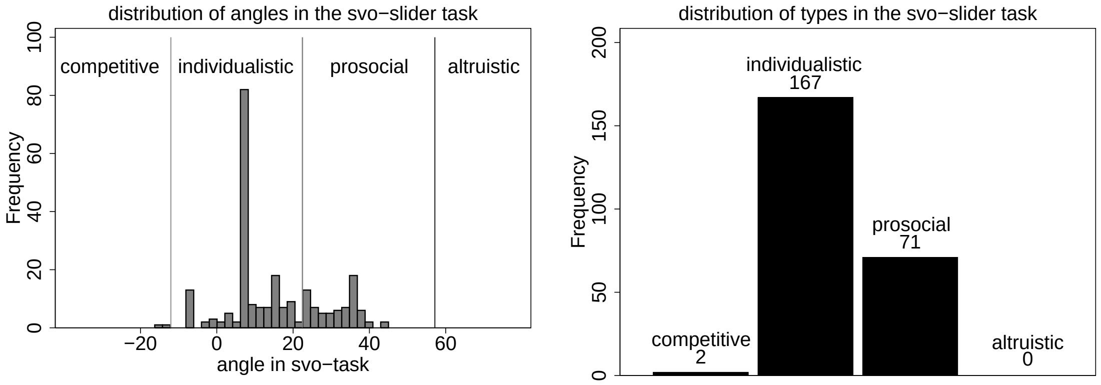

Figure 5. Distribution of angles and types from the svo-slider measure.

7.5. Fairness norms can explain the emerging cooperation norms

In phase two of the experiment, we measured the subjects fairness norms with the svo-slider measure (Murphy et al., 2011). This measure consists of six "'primary items" designed to classify subjects into the four categories *altruistic, prosocial, individualistic* and *competitive.* Nine "secondary items" refine the prosocial types into *joint maximizers*, trying to maximize the joint earnings, and *inequality averse* types, who try to minimize the pay-off differences between two players. The classification algorithm adopted from Murphy et al. (2011) failed to classify many of the secondary items, such that we focuss in our analysis on the primary items only.

In a nutshell, the classification algorithm for the primary items looks for the subjects position on the ring in figure 3 a) to find out the subject's normative type. To do this, we calculate the subject's mean allocation to herself ($\bar{A}_s$) and to the other person ($\bar{A}_o$), and subtract 50 in order to account for the midpoint of the ring being at 50/50. We than determine the resulting angle vector by calculating

$$\text{SVO}^\circ = \arctan\left(\frac{\bar{A}_o - 50}{\bar{A}_s - 50}\right). \tag{8}$$

An altruist would have an angle of greater than 57.15°, a prosocial type an angle between 22.45° and 57.15°, an individualistic type scores between -12.04° and 22.45° and a competitive type below -12.04° (for an extensive discussion of the algorithm as well as for the justification for these boundaries see the corresponding paper by Murphy et al. (2011, p.3)).

Figure 5 reports the distribution of angles and the resulting distribution of types. A vast majority of subjects can be classified as individualistic (167 out of 240 or 69 %), about 85 of them almost exclusively chose the option which maximized their own pay-off. Another 71 (30 %) are classified as prosocial, only 2 (1 %) as competitive and no subject was classified as being altruistic. Due to the small number of competitive players, we excluded them from the following analysis.[9]

Can the types predict the content of the emerging norms? We test our hypotheses 4a-c by estimating a random effects logistic regression with dummies for the types as dependent variables and the respective norms or cooperation as independent variable. In addition to that, we include a dummy to control for the partner's type, taking the value 1 if both partners share the same type.

$$\mathbf{x}_{ij}\beta = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{prosocial} + \boldsymbol{\beta_2}\text{same type} + \nu_\mathbf{j}, \tag{9}$$

The model takes the individualistic type as reference category and estimates the difference in cooperation propensities (first model of table 4) the propensity to establish a turn taking norm (second model) and the propensity to establish a pure strategy norm (third model) for the other remaining types.

Result 4. ***(Fairness norms)***
As compared to individualistic maximizers, the emergence of

a cooperation is ***positively affected*** *by fairness norms.*

[9] We ran the same regressions with and without competititive players and the results are robust.

	sign	cooperation	sign	turn taking	sign	pure strategies
individualistic		ref.		ref.		ref.
prosocial	+	0.932***	+	1.517**	0	0.575
		[0.438,1.426]		[0.401,2.633]		[-0.672,1.823]
partner same type		0.242		1.275*		-1.039
		[-0.400,0.885]		[0.131,2.420]		[-3.227,1.149]
intercept		-2.164***		-6.110***		-6.722***
		[-2.528,-1.800]		[-7.025,-5.195]		[-7.847,-5.596]
random effects						
var(intercept)		1.082***		2.116***		2.093***
		[0.797,1.366]		[1.820,2.412]		[1.683,2.504]
decisions		6050		6050		6050
subjects		238		238		238

bootstrapped 95% confidence intervals in brackets,

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Emergence of cooperation norms as a consequence of different fairness norms. Different forms of cooperation are likely to emerge for different types of actors. All models are random effects logistic regressions controlling for correlated errors on the subject level (see equation 9). The model on the left predicts the general degree of cooperation (i.e. non-D choices) among prosocial and maximizing types. The second model investigates the joint probabilities of turn taking between two choices, given the fairness types. The third model on the right predicts cooperation in pure strategies for the two types. The "sign"-columns list the expected signs as predicted by "behavioral" theory. Coefficients and confidence intervals are based on 1000 bootstraps.

b *turn taking norms is* ***positively affected*** *by fairness norms.*

c *pure strategy norms is* ***not affected*** *by fairness norms.*

As predicted by hypothesis 4a, the first model indicates that prosocial norms increase the probability of choosing a cooperative option. Also result 4b and 4c are in line with our behavioral theory: Prosocial types are better capable of implementing turn taking but do not differ significantly in their propensity to engage in pure strategy norms. Prosocial types have a desire to counterbalance both player's outcomes or to increase the global efficiency, such that they are more likely to engage in turn taking agreements, which does not hold for pure strategy norms.

8. Discussion

This paper theoretically and experimentally studies the emergence of different cooperation norms such as turn taking or pure strategy play in a series of indefinitely repeated normal form games. We study a pay-off symmetric and asymmetric Prisoner's Dilemma (PD) and compare it to a "coordinate-to-cooperate" Battle of the Prisoner's Dilemma (BOPD). The latter can be described as a 3×3 PD where subjects can gain additional surplus if they coordinate in a "Battle of the Sexes" game on how to distribute these newly generated gains of cooperation. We manipulate the asymmetry of the pay-offs and the probability the shadow of the future, i.e. the probability that another instance of the game is played with the same partner.

The games studied here extend the existing literature on cooperation in social dilemmas with exogenous social inequalities by introducing different socially efficient but pay-off discriminating equilibria. The pay-off asymmetry gives rise to a normative conflict about which norm should be adhered to, which is not affected by concerns for efficiency.

The hypothesis derived by means of game theory are quite successful in predicting the general effects of our experimental manipulations: There is more cooperation if the shadow of the future is long and if the opportunity costs of cooperation are low as in the BOPD. Pay-off asymmetry has a positive effect on cooperation in the BOPD, but other than predicted it has a negative effect on cooperation rates in the PD.

However, game theory does not add much to the explanation of *which* norms of behavior are likely to emerge. Most of the hypotheses concerning the emergence of turn taking norms or pure strategy norms are inconsistent with the data. Instead, the type of emerging norms can be quite successfully predicted by the fairness norms held by the interacting players.

We conclude that endogenous fairness norms are an important predictor for the emergence of cooperation norms and the understanding of the equilibrium selection process.

9. Appendix

Lemma 1. (***Equilibria in PD*** *(Axelrod, 1984)*)
Mutual cooperation is a Nash equilibrium in the iterated PD for a pair of TRIGGER strategies and a discount parameter δ if

$$\delta > \frac{T-R}{T-P}$$

Proof. The proof has to show that the gains from mutual cooperation outweigh the gains from one-sided defection followed by eternal defection of the other player. Mutual cooperation returns R for all periods, whereas unilateral defection yields $T > R$ for one period and due to TRIGGER only $P < R$ for the subsequent ones. However, future periods are discounted by δ. Hence, we have to investigate for which δ the following inequality holds:

$$R + \delta R + \delta^2 R + \ldots \geq T + \delta P + \delta^2 P + \ldots$$
$$\frac{R}{1-\delta} > T + \frac{\delta P}{1-\delta}$$

Some straightforward computation gives the critical value of

$$\delta > \frac{T-R}{T-P},$$

for which both players in the PD have no incentive to deviate from C. Consequently, cooperation is an equilibrium in the infinitely repeated PD for sufficiently large δ. □

Lemma 2. (***Equilibria in BOPD***)
Let $\pi_i(\varepsilon)$ be player i's expected pay-off from some combination ε of moves in at least one instance of a super-game of BOPD games. Then ε is a Nash equilibrium in the iterated BOPD for a pair of TRIGGER strategies if*

$$\forall i: \quad \delta > \frac{T-\pi_i}{T-P}.$$

Proof. The proof is equivalent to the case of the prisoner's dilemma. Let $\pi_l(\varepsilon)$ be the the worse-off player's and $\pi_h(\varepsilon)$ the the better-off player's expected per period outcome from some cooperative pattern in *BOPD*, with $T > \pi_h(\varepsilon) \geq \pi_l(\varepsilon)) > P$. We take the minimum of $\pi_i(\varepsilon)$, as the worse-off player has to be better of when adhering to the cooperative equilibrium than switching to defection. Deviating from ε would yield (at best) T for one period, followed by P for the rest of the game such that the expected return is given by

$$U_i(\text{ALL-D}|\text{TRIGGER}^*) = T + \frac{\delta P}{1-\delta},$$

whereas adhering to ε would yield

$$(U_i(\varepsilon|\text{TRIGGER}^*) = \frac{\pi_l(\varepsilon))}{1-\delta}.$$

The pay-off relation

$$U_i(\varepsilon|\text{TRIGGER}^*) > U_i(\text{ALL-D}|\text{TRIGGER}^*)$$

holds for the worse-off player if and only if

$$\frac{\pi_l(\varepsilon)}{1-\delta} > T + \frac{\delta P}{1-\delta}$$
$$\delta > \frac{T-(\pi_l(\varepsilon))}{T-P}.$$

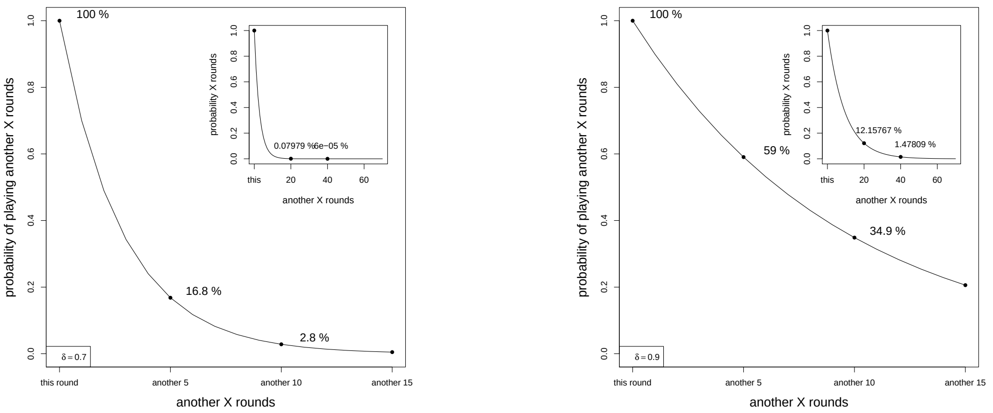

Figure 6. Probabilities of playing future rounds as presented to the experimental subjects.

Since $\pi_h(\varepsilon) \geq \pi_l(\varepsilon)$, neither of the players can gain by deviating from ε if δ is sufficiently large. Hence, ε is an equilibrium in $BOPD$. □

References

Abreu, Dilip, "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica*, 1988, *56* (2), pp. 383–396.

Aksoy, Ozan and Jeroen Weesie, "Inequality and Procedural Justice in Social Dilemmas.," *Journal of Mathematical Sociology*, 2009, *33* (4), 303 – 322.

Axelrod, Robert, *The evolution of cooperation*, New York: Basic Books, 1984.

Bicchieri, Cristina, "Norms of Cooperation," *Ethics*, 1990, *100* (4), 838–861.

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

Bruttel, Lisa V., Werner Güth, and Ulrich Kamecke, "Time to Defect: Repeated Prisoners' Dilemma Experiments with Uncertain Horizon," *Jena Economic Research Papers*, 2007, *2007*, 098.

_ , _ , _ , and Vera Popova, "Voluntary cooperation based on equilibrium retribution: an experiment testing finite-horizon folk theorems," *Jena Economic Research Papers*, 2009, *30* (2009,030).

Buckley, Edward and Rachel Croson, "Income and wealth heterogeneity in the voluntary provision of linear public goods," *Journal of Public Economics*, 2006, *90* (4-5), 935 – 955.

Chan, Kenneth S., Stuart Mestelman, Rob Moir, and R. Andrew Muller, "The Voluntary Provision of Public Goods under Varying Income Distributions," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 1996, *29* (1), pp. 54–69.

Coleman, James Samuel, *Foundations of Social Theory*, Cambridge, Mass. [u.a.]: Belknap Pr., 1990.

de Jasay, Anthony, Werner Güth, Hartmut Kliemt, and Axel Ockenfels, "Take or Leave? Distribution in Asymmetric One-Off Conflict," *Kyklos*, 2004, *57* (2), 217–235.

Efron, Bradley, "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 1987, *82* (397), pp. 171–185.

_ and Rob Tibshirani, *An introduction to the bootstrap*, Vol. 57, Chapman & Hall/CRC, 1993.

Ellickson, Robert C., *Order without law*, Cambridge, Mass.: Harvard University Press, 1991.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

Fischbacher, Urs, "Ztree- Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 2007, *10* (2), 334–352.

Fudenberg, Drew and Eric Maskin, "The folk theorem for repeated games with discounting and incomplete information," *Econometrica*, 1986, *54*, 533–554.

Glöckner, Andreas, Bernd Irlenbusch, Sebastian Kube, Andreas Nicklisch, and Hans-Theo Normann, "Leading with(out) Sacrifice? A Public-Goods Experiment with a Priviledged Player," *Economic Inquiry*, 2011, *49* (2), 591–597.

Gonzalez, Luis G., Werner Güth, and M.Vittoria Levati, "When does the game end? Public goods experiments with non-definite and non-commonly known time horizons," *Economics Letters*, 2005, *88* (2), 221–226.

Gouldner, Alvin W., "The Norm of Reciprocity: A Preliminary Statement," *American Sociological Review*, 1960, *25* (2), pp. 161–178.

Greiner, Ben, "An Online Recruitment System for Economic Experiments," in Kurt Kremer and Volker Macho, eds., *GWDG Bericht 63*, Göttingen: Ges. für Wiss. Datenverarbeitung, 2004, pp. 79–93.

Kroll, Stephan, Todd Cherry, and Jason Shogren, "The impact of endowment heterogeneity and origin on contributions in best-shot public good games," *Experimental Economics*, 2007, *10* (4), 411–428.

Liebrand, Wim B. G. and Charles G. McClintock, "The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation," *European Journal of Personality*, 1988, *2* (3), 217–230.

Luce, R. Duncan and Howard Raiffa, *Games and Decision. Introduction and Critical Survey*, New York: Dover, 1989.

Marwell, Gerald and Ruth E. Ames, "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem," *The American Journal of Sociology*, 1979, *84* (6), 1335–1360.

Miller, Luis, Heiko Rauhut, and Fabian Winter, "The emergence of norms from conflicts over just distributions," *Jena Economic Research Papers*, 2011, *2011-18.*

Murphy, Ryan O., Kurt Ackermann, and Michel J.J. Handgraaf, "Measuring Social Value Orientation," *Working paper, ETH Zrich, Chair of Decision Theory and Behavioral Game Theory*, 2011.

Nikiforakis, N., C.N. Noussair, and T. Wilkening, "Normative Conflict & Feuds: The Limits of Self-Enforcement," *Department of Economics-Working Papers Series*, 2011.

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants With and Without a Sword: Self-Governance is Possible," *The American Political Science Review*, 1992, *86* (2), 404–417.

Rabe-Hesketh, Sophia and Anders Skrondal, *Multilevel and longitudinal modeling using stata*, College Station, Tex: Stata Press, 2005.

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

Rauhut, Heiko and Fabian Winter, "A Sociological Perspective on Measuring Social Norms by Means of Strategy Method Experiments," *Social Science Research*, 2010, *39* (6), 1181 – 1194.

Reuben, Ernsesto and Arno Riedl, "Enforcement of contribution norms in public good games with heterogeneous populations," *IZA Discussion Papers*, 2011.

Schmelz, Katrin, "E-nstructions: Using Electronic Instructions in Laboratory Experiments," *Jena Economic Research Papers*, 2010.

Snijders, Tom A.B. and Roel Bosker, *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 1 ed., Sage Publications Ltd, 12 1999.

Taylor, Michael, *Anarchy and cooperation*, New York: John Wiley and Sons, 1976.

Tutic, Andreas and Ulf Liebe, "A Theory of Status-Mediated Inequity Aversion," *The Journal of Mathematical Sociology*, 2009, *33* (3), 157–195.

Ullmann-Margalit, Edna, *The emergence of norms* Clarendon library of logic and philosophy, Oxford: Clarendon Press, 1977.

van Dijk, Eric and Henk Wilke, "Coordination Rules in Asymmetric Social Dilemmas: A Comparison between Public Good Dilemmas and Resource Dilemmas," *Journal of Experimental Social Psychology*, 1995, *31* (1), 1 – 27.

Vogt, Sonja and Jeroen Weesie, "Social support among heterogeneous partners," *Analyse & Kritik*, 2004, *2*, 398–422.

Voss, Thomas, "Game-Theoretical Perspectives on the Emergence of Social Norms," in Michael Hechter and Karl-Dieter Opp, eds., *Social norms*, New York: Rusell Sage Foundation, 2001.

Winter, Fabian, Heiko Rauhut, and Dirk Helbing, "How norms can generate conflict: An experiment on the failure of cooperative micro-motives on the macro-level," *Social Forces*, forthcoming.

Yamagishi, Toshio, "The Provision of a sanctioning system as a Public Good," *Journal of Personality & Social Psychology*, Jul 1986, *51* (1), 110–116.