Social Preferences Under the Shadow of the Future*

Felix Kölle† — *University of Cologne*

Simone Quercia‡ — *University of Verona*

Egon Tripodi§ — *European University Institute & University of Bonn*

June 2020

Abstract

Social interactions predominantly take place under the shadow of the future. Previous literature on infinitely repeated games has highlighted the primary role of self-interested strategic considerations in explaining outcomes. Using indefinitely repeated prisoner's dilemma games, this paper demonstrates experimentally the importance of social preferences for achieving efficient cooperative outcomes. Sorting agents by their prosociality, we find that cooperation is three to four times higher among prosocial players compared to selfish players. We also show that social preferences are less important when individuals interact in mixed populations. This can explain why the influence of social preferences has not been detected in previous studies.

Keywords: cooperation, infinitely repeated game, prisoner's dilemma, social preferences, experiment.

JEL Classification: C73, C91, C92

*The project was approved by the Ethics Committee at the University of Cologne. Both our main experiment (#18979) and our follow-up (#28887) were pre-registered on the AsPredicted platform. Research funding from the Reinhardt Selten Institute is gratefully acknowledged. We are especially grateful to Maria Bigoni, Guillaume Fréchette, David K. Levine, and Daniele Nosenzo for very constructive feedback. We thank Carina Lenze, Margarita Radkova, Vincent Selz and Valerie Stottuth for their help in running the experiment, and Jae Youn Nam and Luis Wardenbach for helping program the experimental software. All errors remain our own.

†University of Cologne, Albertus Magnus Platz, 50923 Cologne, Email: felix.koelle@uni-koeln.de.

‡University of Verona, Via Cantarane 24, 37129 Verona, Email: simone.quercia@univr.it

§European University Institute and University of Bonn, Adenauerallee 24-42, 53113 Bonn, Email: egon.tripodi@eui.eu.

1 Introduction

Social dilemmas are ubiquitous in nature and exist at all levels of human society, ranging from team production and collaborations among firms to the maintenance of natural resources and the provision of public goods. The unifying element behind these examples is the fundamental tension between individual and collective interest. The simplest and most commonly used game to study this tension is the prisoner's dilemma, which has raised continued interest across the social sciences. In economics, the theory of infinitely repeated games has identified conditions under which cooperation can thrive. In particular, through the punishment of opportunistic behavior (and reward of cooperation), according to the Folk theorem, cooperation can be sustained in equilibrium if agents are sufficiently patient (Fudenberg and Maskin, 1986). While a series of experiments has provided empirical support for this theory, a substantial amount of unexplained variation remains (see Dal Bó and Fréchette, 2018, for a recent overview).

The theory of infinitely repeated games is built on the assumption that decision makers are exclusively motivated by their own material benefit. As a consequence, cooperation can only arise due to strategic motivations, i.e., the prospects of future interactions. However, there is by now a large literature providing evidence from a wide variety of contexts that many individuals are not only motivated by their own material payoffs, but that they also care about the well-being of others. Formal models of such other-regarding concerns have been quite successful in explaining several patterns of behavior observed in laboratory experiments and in the field (see Sobel, 2005; Fehr and Schmidt, 2006; Cooper and Kagel, 2016, for overviews of the literature). In the context of cooperation, previous studies have shown that findings from one-shot and finitely repeated games can be well explained by the fact that individuals are heterogeneous with regard to their taste for (conditional) cooperation (Gintis, 2000; Fischbacher et al., 2001; Fehr and Fischbacher, 2003; Fischbacher and Gächter, 2010; Fehr and Schurtenberger, 2018). Yet, despite the importance of social preferences in explaining behavior in these and related contexts, so far, they have been thought to play only a minor role in infinitely repeated games, as recently noted by Dal Bó and Fréchette (2018, p. 88): "It is interesting that altruistic and trusting tendencies (as captured by the dictator and trust games) do not seem to play an important role in infinitely repeated games."

In this paper, we provide a clean test of the importance of social preferences in explaining cooperation in infinitely repeated games, and offer a new rationale for why the role of social preferences has remained undetected so far. Our empirical strategy is based on a laboratory experiment, in which we first elicit players' revealed preferences for cooperation in a se-

quential one-shot version of the prisoner's dilemma game using the strategy method (Selten, 1967; Fischbacher et al., 2001). This method allows us to distinguish between individuals who are predominantly concerned by their own material benefit (i.e., those who choose to defect when the other player cooperates), and individuals who display a sufficiently strong degree of social preferences as revealed by their willingness to cooperate (rather than defect) in case the other player cooperates. In the following, we will call these individuals *selfish* and *prosocial*, respectively.

Using these elicited preferences, in the second part of our experiment, we sort players into different groups and let them play twenty supergames of an indefinitely version of the game using a random termination rule (Roth and Murnighan, 1978). The groups thereby differ with regard to the composition of player types. We consider three types of groups: (i) *selfish groups* that only consist of participants classified as selfish, (ii) *prosocial groups* that only consist of participants classified as as prosocial, and (iii) *mixed groups* that consist of a combination of both prosocial and selfish types. This latter group is most comparable to the ones studied in previous literature, which has matched players at random and has recruited participants from subject pools that are similar to ours. Importantly, in order to have sharp theoretical predictions with regard to difference in expected cooperation between selfish and prosocial groups, before the start of the game, players are informed about the type of group they are interacting in. Furthermore, we set the continuation probability (δ) in the repeated game such that full defection is the only equilibrium outcome among purely self-interested players. For prosocial groups, in contrast, we show that under the condition of segregated groups and common knowledge thereof, mutual cooperation can (on top of mutual defection) be sustained in equilibrium irrespective of the level of δ.

The results of our experiment reveal that, in line with the theoretical predictions, groups consisting of selfish types achieve very low levels of cooperation, especially after subjects gained some experience in the game. Mixed groups start off with higher cooperation rates, but converge to very similar levels as selfish groups. This is consistent with a large experimental literature showing that cooperation quickly converges to very low levels when mutual cooperation is not an equilibrium (Dal Bó and Fréchette, 2018). In stark contrast to that, prosocial groups manage to coordinate on very high levels of cooperation throughout the entire game. Averaged over all twenty supergames, cooperation rates amount to 72% in prosocial groups, compared to 35% in mixed groups and 18% in selfish groups. Furthermore, when analyzing the behavior of prosocial and selfish players in mixed groups, we find that they cooperate to a very similar extent. These results demonstrate that while social pref-

erences have a strong promoting effect on cooperation when players interact in segregated groups, this effect is reduced when interacting in mixed populations.[1]

Further support for this finding comes from the analysis of the prevalence of specific repeated game strategies among selfish and prosocial players. For selfish players we find Always Defect (AD) to be by far the most common strategy, irrespective of the type of group they are interacting in. For prosocial players, in contrast, we find that the vast majority chooses cooperative strategies (with Tit-for-Tat being the modal one), but only when paired among themselves. When playing together with selfish types in mixed groups, instead, the most prominent strategy is AD. Taken together, these results can explain why previous literature has not detected a strong impact of social preferences in infinitely repeated games, as they typically have grouped subjects at random, thus creating mixed groups.

In a follow-up experiment, we further show that the power of social preferences in increasing cooperation carries over to a situation in which cooperation can be sustained in equilibrium even among purely self-interested players ($\delta > \delta^{SPE}$). Specifically, while we find that all group types react to the increase in the continuation probability with increasing cooperation rates, the gap in cooperation between selfish groups (40%) and prosocial groups (85%) remains effectively unaltered. This suggests that social preferences can serve as an important equilibrium selection device when strategic incentives are such that both cooperation and defection can be part of an equilibrium.

Our approach relies on the ability to classify individuals based on their degree of social preferences. To test the validity of our approach, we show that the types we elicit in the first part of our experiment are related to a series of individual characteristics that previous literature has associated with (preferences for) cooperation. In particular, we find a strong association between our measure of social preferences and an incentivized measure of norm-following (Kimbrough and Vostroknutov, 2018), supporting the notion that conditional cooperation is related to norm adherence (Fehr and Schurtenberger, 2018). Furthermore, in line with previous evidence we find that prosocial and selfish players differ along important personality dimensions such as agreeableness and conscientiousness (Volk et al., 2012; Proto et al., 2019). Finally, we show that other factors such as risk preferences and intelligence (as measured by a ten-item Raven matrices test) are unrelated to our social preference measure,

[1] The results on selfish and prosocial groups are reminiscent of findings from previous studies that have investigated the role of sorting in finitely repeated games. For example, Gächter and Thöni (2005) show that subjects matched according to a one-shot contribution in a public goods game achieve substantially different levels of cooperation in a subsequent ten-period public goods game. A similar result has been found by Kimbrough and Vostroknutov (2016) who sort subjects according to their willingness to follow rules.

suggesting that differences in cognitive abilities or confusion are unlikely explanations for our findings.

Our paper contributes to the growing experimental literature on indefinitely repeated games studying the conditions that favor the emergence of cooperation (see e.g., Palfrey and Rosenthal, 1994; Dal Bó, 2005; Engle-Warnick and Slonim, 2006; Aoyagi and Fréchette, 2009; Camera and Casari, 2009; Fudenberg et al., 2012; Bigoni et al., 2015; Arechar et al., 2017; Fréchette and Yuksel, 2017; Aoyagi et al., 2019). In particular, previous studies have shown that while cooperation is typically higher when it can be supported in equilibrium compared to when it cannot, subjects often fail to make the best out of it. Furthermore, while coordination failure becomes less likely when cooperation is also risk dominant (Blonski et al., 2011; Dal Bó and Fréchette, 2011), even in this case substantial variation remains. This has led researchers to start looking into additional factors that can explain the variation in cooperation outcomes (see, e.g., Proto et al., 2019, for a recent effort to investigate the influence of intelligence and personality).

Two previous studies (Dreber et al., 2014; Davis et al., 2016) have investigated the role of social preferences in infinitely repeated games by correlating behavior in the repeated game with donation decisions elicited ex-post via dictator games. The evidence provided in these papers is mixed. Based on their results, Dreber et al. (2014) conclude that social preferences are not a key determinant of cooperation in repeated games. Using a different approach, Reuben and Suetens (2012) disentangle strategic from non-strategic motivations for cooperation by letting subjects condition their choice on whether the current round of an interaction is the last round or not. They find that most of the cooperation in the repeated prisoner's dilemma game is strategically motivated. Most closely related to our paper is a recent study by Kartal and Müller (2018) who propose a model with private information in which players are heterogeneous with regard to their taste for cooperation. They show that when the continuation probability of the repeated game is not conducive to cooperation, assuming heterogeneity in social preferences can introduce Bayesian Nash equilibria, in which people with strong enough social preferences play cooperative strategies and other players defect. They further show that in this case, reducing the strategic risk by letting players move sequentially rather than simultaneously should increase cooperation. This prediction is borne out by the data of their experiment, a finding that has also been reported by Ghidoni and Suetens (2019).

Our approach differs from the ones applied in previous papers as follows. In contrast to Kartal and Müller (2018) and Ghidoni and Suetens (2019) who infer the role of other-

regarding concerns in repeated games only indirectly by comparing simultaneous with sequential game play, we provide a more direct test by measuring social preferences explicitely and, subsequently, sorting participants in either mixed or segregated groups. Furthermore, in contrast to Dreber et al. (2014) and Davis et al. (2016), we elicit subjects' other-regarding concerns in a task that is more tightly connected to the strategic decision situation participants face in the repeated game. This is important because previous studies have shown that other-regarding motives are highly context-dependent (see Galizzi and Navarro-Martínez, 2019, for recent evidence and a meta-analysis). In addition to the fact that previous studies have only considered mixed groups, this can explain why the correlations between measures of social preferences and repeated game behavior have been found to be low. We circumvent this problem by using the same stage game of the prisoner's dilemma in both parts of the experiment. Together with our design feature of sorting subjects into mixed or segregated groups, this has the major advantage that it allows us to derive tight theoretical predictions for subjects' behavior in the repeated game. As such, our study provides novel insights about the importance of social preferences in explaining cooperation outcomes in repeated games.

The remainder of the paper is structured as follows. Section 2 discusses the theoretical implications of social preferences in the repeated Prisoner's dilemma game. Section 3 describes the experimental design and hypotheses. Section 4, 5 and 6 report our experimental results. Section 7 concludes.

2 Theoretical Considerations

Consider the stage game represented in normal form in panel (a) of Table 1. If $T > R > P > S$, the payoff matrix represents the prisoner's dilemma game. Assuming self-interest and rationality, *Defect* constitutes a dominant strategy and, thus, (*Defect, Defect*) is the unique Nash Equilibrium (NE) in the stage game. Through the logic of backward induction, the same prediction holds when the game is repeated a *finite* number of times.

Table 1: Prisoner's dilemma game

(a) Payoffs

		Player 2	
		Cooperate	*Defect*
Player 1	*Cooperate*	R, R	S, T
	Defect	T, S	P, P

(b) Utilities

		Player 2	
		Cooperate	*Defect*
Player 1	*Cooperate*	$U(R, R)$	$U(S, T)$
	Defect	$U(T, S)$	$U(P, P)$

In *infinitely repeated* contexts, in contrast, the Folk Theorem (Fudenberg and Maskin, 1986) predicts that if agents are sufficiently patient, (*Cooperate, Cooperate*) can be supported as an equilibrium outcome even among completely self-interested individuals. In particular, if both players follow the grim trigger strategy, i.e., start with cooperation in round 1, continue to cooperate until the other player defects, and then defect forever, choosing grim trigger yields a higher payoff than always defect if:

$$\sum_{t=0}^{\infty} \delta^t R > T + \sum_{t=1}^{\infty} \delta^t P$$

or, rearranging,

$$\delta > \hat{\delta}^{SPE} = \frac{T-R}{T-P}.$$

Hence, if players are only interested in their own payoff, mutual cooperation can only be sustained in a subgame perfect equilibrium (SPE) if the shadow of the future is sufficiently long, i.e., if $\delta > \hat{\delta}^{SPE}$.

There is now, however, vast evidence from a variety of contexts that many people are not only motivated by their own payoffs, but that they also care about the well-being of others (see Sobel, 2005; Fehr and Schmidt, 2006; Cooper and Kagel, 2016, for overviews of the literature). For example, previous literature has shown that many people are willing to cooperate even in one-shot prisoner's dilemma games (Mengel, 2017). Several models of social preferences have been proposed to reconcile such behavior (see Rabin, 1993; Levine, 1998; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002, among others). Applying social preferences to the game above affects the mapping of monetary payoffs into utilities leading to the transformed game as displayed in Panel (b) of Table 1.

While in the one-shot game purely self-interested agents have a dominant strategy to defect, i.e., $U(Defect, Cooperate) > U(Cooperate, Cooperate)$ and $U(Defect, Defect) > U(Cooperate, Defect)$, an agent with sufficiently strong social preferences may prefer $U(Cooperate, Cooperate)$ over $U(Defect, Cooperate)$, for example due to inequity concerns (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), feelings of guilt (Battigalli and Dufwenberg, 2007), preference for conforming to others (Bernheim, 1994; Götte and Tripodi, 2018), or the willingness to reward kind actions of others (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). If both players have such preferences, and this is commonly known, the prisoner's dilemma game (in payoffs) turns into a coordination game (in utilities) with multiple equilibria. Thus, under the assumption of common

knowledge of (sufficiently strong) social preferences, mutual cooperation is (on top of mutual defection) a possible equilibrium outcome even of the stage game. As a consequence, mutual cooperation will also be an equilibrium outcome in *finitely* as well as in *infinitely repeated* games.[2]

In our empirical strategy, we will abstract from identifying the exact type of social preferences at play, but instead elicit a revealed preference measure for (conditional) cooperation. From a utility point of view, these preferences can be represented with a general utility function $u_i(R, T, P, S, \beta_i)$ where β_i is the parameter governing social preferences. Our design will reveal which participants display a β_i high enough such that the prisoner's dilemma game (in payoffs) turns into a coordination game (in utilities), and which participants have a sufficiently low β_i such that defection remains the dominant strategy. The advantage of this approach is that because we remain agnostic about the exact social preference motive at play — which likely is different across individuals — we don't have to assume any specific functional form of the utility function.

3 The Experiment

Our experiment consists of three parts. In the first two parts, subjects play different variants of the prisoner's dilemma game as displayed in Table 2. We set the payoff for mutual cooperation to $R = 15$, the payoff for mutual defection to $P = 10$, the temptation payoff to $T = 25$, and the sucker payoff to $S = 0$. In part 1, we elicit a proxy for individuals' social preferences using the strategy method (Selten, 1967). In part 2, participants play an indefinitely repeated version of the same stage game. In part 3, we elicit a series of individual characteristics to assess the validity of our proxy for social preferences. In the following, we explain each part in detail.

3.1 Experimental Design

Part 1: Eliciting preferences for cooperation. In part 1 of the experiment, subjects play a sequential one-shot version of the prisoner's dilemma game as shown in Table 2.

[2] Under some circumstances, e.g., if altruistic preferences or concerns for efficiency become sufficiently strong (Charness and Rabin, 2002), cooperation may even become the dominant strategy. However, previous evidence reveals that the majority of people are willing to cooperate only conditionally on others doing so too, rather than unconditionally (Fischbacher et al., 2001; Chaudhuri, 2011; Gächter et al., 2017), a finding we replicate in our experiment (see Section 4). As a consequence, both mutual cooperation and mutual defection are possible equilibrium outcomes.

Table 2: Payoffs in the prisoner's dilemma game

		Player 2	
		Cooperate	*Defect*
Player 1	*Cooperate*	15, 15	0, 25
	Defect	25, 0	10, 10

Based on the design by Fischbacher et al. (2001), we elicit an individual's willingness to cooperate as a function of the other player's action. To this end, subjects are asked to make one *unconditional* and two *conditional* decisions. In the unconditional (first-mover) decision, subjects are simply asked to choose one of the two options, cooperate or defect. In the conditional (second-mover) decisions, subjects are asked to make a decision contingent on the other player's unconditional decision. Using the strategy method (Selten, 1967), we ask them (i) whether they want to cooperate or defect in case the other player defects, and (ii) whether they want to cooperate or defect in case the other player cooperates. To guarantee incentive compatibility of all choices, at the end of the experiment, in each pair, a random mechanism selects one player as the first-mover and the other player as the second-mover. For the first-mover it is the unconditional decision that is payoff-relevant. For the second-mover one of the two conditional decisions is payoff-relevant. Which one of the two is determined by the first-mover's unconditional decision.

We use a participant's responses in the conditional decisions as a proxy for their social preferences. Previous studies have shown that there is pronounced heterogeneity in individuals' preferences for cooperation. The vast majority of individuals can be classified into one out of two types: free-riders who choose to defect irrespective of the other player's decision, and conditional cooperators who are willing to cooperate if others do so too (Fischbacher et al., 2001; Chaudhuri, 2011; Gächter et al., 2017). While these studies have typically used the entire strategy profile to classify individuals, in light of our discussion in Section 2, for our purpose it is sufficient to only consider a subject's revealed preference when responding to other's cooperation. This is because the condition that guarantees the existence of a mutual cooperation equilibrium is that individuals prefer to cooperate when their matched counterpart cooperates. Therefore, we classify an individual as prosocial if she responds to the other's choice to cooperate with cooperation, and as selfish if she responds to other's cooperation with defection.

Part 2: Indefinitely repeated game. In part 2 of the experiment, participants play the same stage game as in part 1 for an indefinite number of times. We use a random continuation rule: after each round, given a fixed and known continuation probability δ, a random device determines whether the game goes on for another round or stops. We fix the continuation probability at $\delta = 0.6 < \hat{\delta}^{SPE} = 0.67$, such that under narrow self-interest (and knowledge thereof), mutual cooperation cannot be supported as an equilibrium outcome in the repeated game. Every time the game stops, a supergame ends and a new one begins. Subjects play a total of twenty supergames and this was announced at the beginning of part 2. Subjects remain matched with the same counterpart for all rounds within a supergame, but are randomly re-matched with a new counterpart at the beginning of a new supergame.

The crucial feature of our experimental design is that we manipulate the composition of the groups that subjects are assigned to. Within each session, we create three different types of groups, with $n = 10$ subjects each: prosocial groups, selfish groups, and mixed groups. Groups differ with regard to the composition of types as determined in part 1 of the experiment. Specifically, prosocial groups consist only of subjects who in their conditional decision chose to cooperate if the other player cooperates. Selfish groups, in contrast, consist only of subjects who chose to defect if the other player cooperates. Finally, mixed groups consist of a combination of both these types.

Importantly, subjects were informed about the details of part 2 and the type of group they are assigned to only before the start of the repeated game in part 2. Not revealing detailed information about part 2 before subjects completed part 1 is important in order to elicit a clean measure of subjects' social preferences that is not biased by strategic considerations (see Proto et al., 2019, for a similar approach). At the beginning of Part 2, we communicated the exact choice that was used to generate the matching, that is, in prosocial (selfish) groups we explained that all subjects in the groups chose to *cooperate* (*defect*) in the conditional decision in case their counterpart chose to *cooperate*. For mixed groups, we explained that the group was composed by some participants who had chosen to *cooperate* and some who had chosen to *defect* in response to other's cooperation.[3] Providing this information before

[3] A critical issue when performing experiments where subjects are matched according to their behavior in previous parts of the experiment is the possibility of deceiving subjects by withholding potentially payoff-relevant information from them. To tackle this issue, at the beginning of the experiment we informed subjects that we would use their part 1 decisions for part 2. In particular, we told subjects in the instructions of part 1 that in part 2, they would either interact with players who in part 1 made the same or different choices than themselves. This statement was true irrespective of the own type, as subjects were always either placed in a segregated group or in a mixed group. Given the uncertainty on the group matching and the fact that we pay only one out of the two parts at random (see below), it is unlikely that individuals distorted their choices in part 1 to affect the matching of part 2. It is further worth noting that if strategic considerations

the start of part 2 is crucial because, in light of the theoretical considerations laid out in Section 2, an equilibrium with mutual cooperation exists if (i) all players have sufficiently strong social preferences, and (ii) this is common knowledge among all players. Hence, in order to test for the importance of social preferences in repeated games, both the creation of segregated groups and providing information about the group composition are essential design features.[4]

Another crucial feature of our design is that we are eliciting subjects' other-regarding concerns in a task that is tightly connected to the strategic decision situation participants face in the repeated game. This is important because previous studies have shown that other-regarding motives are highly context-dependent (see e.g., Galizzi and Navarro-Martínez, 2019). Together with the fact that previous studies have only investigated behavior in mixed groups, this can explain why they have found low correlations between measures of social preference such as those from dictator and trust games, and repeated game behavior (see Dreber et al. (2014); Davis et al. (2016)). We circumvent this problem by using the same stage game of the prisoner's dilemma in parts 1 and 2.

Part 3: Eliciting individual characteristics. Given recent evidence highlighting the importance of personal characteristics for outcomes in repeated interactions (Proto et al., 2019), and in order to provide insights into the "behavioral validity" of our social preference measure, in part 3 of our experiment we elicit a series of individual characteristics that have been previously associated with cooperation. In particular, given the importance of personality and intelligence (Proto et al., 2019), we elicit a big-five personality inventory (Schupp and Gerlitz, 2008) and a measure of IQ using a ten-item Raven's progressive matrices test (Raven, 2000). We further elicit subjects' general risk attitudes (Dohmen et al., 2011) and gender. Lastly, we elicit subjects' propensity to follow norms. This is particularly interesting in our setting, as a recent study by Fehr and Schurtenberger (2018) has argued that (conditional) cooperation is associated with norm following and normative behavior. To

would have a strong effect on subjects' responses in part 1, e.g. if many selfish types would pretend to be cooperative in order to be matched with and exploit prosocial types in part 2, this should reduce potential differences between our two segregated groups and, thus, work against our hypothesized effect.

[4] Note that our treatments thus might not only manipulate the composition of types within groups, but also may shift players' expectations on the likelihood that other group members will cooperate, thereby creating a focal point. Controlling for this effect would require to communicate group composition without shifting beliefs, which seems hard in our design. Notice, however, that beliefs alone can not explain any differences in cooperation outcomes between our segregated groups. The reason is that for $\delta < \hat{\delta}^{SPE}$, the basin of attraction is equal to 1, i.e., for a completely self-interested agent, always defect is optimal regardless of beliefs about the other player's behavior. Hence, social preferences are necessary to explain potential differences in cooperation.

test for this, we implement an incentivized norm-following task as introduced by Kimbrough and Vostroknutov (2018). In this task, participants are asked to allocate 50 balls between two urns. The blue urn pays 0.02 Euro per ball placed and the yellow urn pays 0.04 Euro per ball. The instructions specify that "the rule is to place the balls in the blue urn" (see Appendix D.1 for the instructions and Appendix D.2 for a visualization). As shown by Kimbrough and Vostroknutov (2018), the number of balls placed in the blue urn constitutes a proxy for individuals' propensity to follow norms.

3.2 Hypotheses

We start our discussion on the expected levels of cooperation with selfish groups. Recall that in these groups, we match together only individuals who, in part 1 of our experiment, have revealed a preference for defection when the other player cooperates and, thus, a sufficiently low degree of social preferences (β_i). As discussed in Section 2, when $\delta < \hat{\delta}^{SPE}$, the unique equilibrium outcome among such (mainly self-interested) individuals is full defection. As a result, we expect very low levels of cooperation in these groups, especially as subjects gain experience in the game. This leads us to our first hypothesis:

Hypothesis 1. *Selfish groups converge to full defection over time.*

Next, we consider the predictions for prosocial groups. As discussed in Section 2, the conditions for the existence of cooperative equilibria in these groups are that (i) subjects prefer to cooperate rather than defect when the other player cooperates and (ii) everyone knows that all group members have such preferences. Our experimental design guarantees that these two conditions are satisfied. Yet, even in this case both mutual cooperation and mutual defection are feasible equilibrium outcomes and, hence, it is a priori unclear at which level of cooperation prosocial groups will coordinate on. We hypothesize that social preferences may serve as an equilibrium selection device in order to select the Pareto-efficient equilibrium. This can happen both because our treatment enhances knowledge on the preferences of other group members, and because the information provided can help coordinate expectations about the behavior of others. This leads us to our second hypothesis:

Hypothesis 2. *Prosocial groups are able to achieve higher levels of cooperation than selfish groups.*

Finally, we turn to the prediction for mixed groups, in which some members exhibit a strong degree of social preferences while others are mainly self-interested. As recently shown

theoretically by Kartal and Müller (2018), assuming that participants know the distribution of types in the population, in mixed groups Bayesian Nash equilibria exist where, depending on their degree of social preferences, some players play cooperative strategies (such as Grim Trigger or Tit-for-Tat) and some other players always defect. These equilibria are in addition to the full defection equilibrium. Hence, the set of possible equilibria in mixed groups is larger and includes more cooperative equilibria than in selfish groups. Compared to prosocial groups, in contrast, the set of equilibria is less cooperative as mutual cooperation is not feasible.[5] Additional insights regarding our mixed groups can be derived from previous studies on indefinitely repeated games, which have paired subjects at random and have recruited participants from student subject pools similar to ours. Most of these studies have found that when $\delta < \hat{\delta}^{SPE}$ cooperation reaches very low levels, especially after some time (Dal Bó and Fréchette, 2018). This leads to our third hypothesis:

Hypothesis 3. *Cooperation in mixed groups will be weakly higher than in selfish groups but lower than in prosocial groups.*

3.3 Procedures

All sessions of our experiment were conducted at the Cologne Laboratory for Economic Research (CLER). In total, we recruited $n = 240$ participants in nine equally sized sessions of 30 participants each. The experimental software was programmed using oTree (Chen et al., 2016) and student participants from various disciplines were recruited using ORSEE (Greiner, 2015). The experiment was pre-registered on the AsPredicted platform (#18979).

At the beginning of the experiment, subjects were informed about the three-part nature of the experiment. They then received instructions explaining the general decision situation of the prisoner's dilemma game including some examples (see Appendix C for an English copy of the instructions). After that, subjects read the instructions for part 1 of the experiment, followed by control questions designed to ensure subjects' understanding of the game. Only after all participants answered all questions correctly, part 1 started. Upon completion of part 1, subjects received instructions for part 2. The instructions were again followed by a set

[5] The equilibria derived in Kartal and Müller (2018) depend on a cutoff parameter such that individuals with weak social preferences (below the cutoff) play always defect and players with strong social preferences (above the cutoff) play cooperative strategies. As we do not derive the cutoff as in Kartal and Müller (2018), some of the equilibria they characterize may be played also in our selfish groups. However, as mixed groups are formed by individuals who are more prosocial on average than individuals in the selfish groups, according to Kartal and Müller (2018) the set of equilibria in mixed groups includes more cooperative equilibria than in selfish groups.

of control question, testing subjects' understanding of the matching procedure, payments, and the game structure. After answering all questions correctly, subjects were then informed about the type of group they were assigned to. In mixed groups, subjects were told that they will interact with both prosocial and selfish players, but nothing was said about the relative frequency of the respective types. After that, part 2 started. The length of each supergame was determined randomly within each session. That implies that the total number of rounds played is the same across all matching groups within a session, but differs across sessions. The length of the supergames ranged between 1 and 10 rounds, with an average (sd) of 2.32 (1.78).

After finishing part 2, but before learning their earnings from the experiment, subjects were introduced to part 3, containing the norm-following task, the IQ test, the big-five personality inventory, the risk question, and demographic questions. At the end of part 3, we randomly selected either part 1 or part 2 (with equal chance) to determine subjects' earnings. If part 1 was selected, subjects were paid either according to their unconditional or their conditional decision as described above. If part 2 was selected, the computer randomly selected one out of the twenty supergames and paid the last round of that supergame (see Dal Bó and Fréchette, 2018, for a discussion of different payment methods). We chose to pay only one of the two parts in order to avoid spillover effects due to, e.g., income effects, and to limit strategic incentives for subjects to distort their choices in part 1. In addition, subjects received their earnings from the norm-following task, ranging between 1 and 2 Euro. On average, subjects earned 17.70 Euros for sessions that lasted around one hour.

4 Results

We organize the discussion of our results as follows. We start by describing the results of the strategy method in part 1 of our experiment. We then show how grouping subjects with different degrees of social preferences into mixed or segregated groups affects cooperation in the indefinitely repeated game in part 2. After that, we use the Structural Frequency Estimation Method (SFEM) developed by Dal Bó and Fréchette (2011) to estimate repeated-game strategies used in the different groups.

4.1 Cooperation types

In the unconditional decision of part 1 of the experiment, 35% of the participants chose to cooperate. In the conditional decisions, cooperation drops to 8% when choosing conditional on the partner's choice to defect, and increases to 42% when choosing conditional on the partner's choice to cooperate. Based on the latter choice, we classify subjects either as prosocial or selfish, depending on whether they respond to their partner's unconditional choice to cooperate with cooperation or defection, respectively. We thus have 42% prosocial types and 58% selfish types (see Table A1 for a full breakdown of choices). Note that given that almost all of our subjects (92%) chose to defect conditional on other's defection, we find that the large majority of subjects (88%) which we classify as prosocial are conditional cooperators, i.e., they are willing to cooperate only if the other player does so too, and that almost all subjects (96%) which we classify as selfish are free-riders who choose to defect irrespective of the other player's choice.[6] Using this classification, we then formed the different groups as described above. In total we had seven groups consisting of only prosocial types (prosocial groups), eleven groups consisting of only selfish types (selfish groups), and six groups consisting of both prosocial and selfish types (mixed groups). In the latter, the average share of prosocial and selfish types was equal to 50%.[7]

4.2 The effects of group composition on cooperation rates

Figure 1 illustrates for each group type how aggregate cooperation rates evolve across the twenty supergames.[8] The left panel shows first round cooperation rates and the right panel shows cooperation rates from all rounds. Focusing on cooperation rates based on first round decisions is informative, because supergames may have different lengths, and cooperation rates may depend on histories within these supergames. For both measures, we observe pronounced differences in cooperation rates across the different types of groups. In particular, in line with Hypothesis 2, we observe much higher cooperation rates in prosocial compared to selfish groups. Furthermore, while in prosocial groups we find very high levels of cooper-

[6] When following previous studies by using both conditional choices in order to classify people into types, we find 5% unconditional cooperators, 37% conditional cooperators, 56% unconditional defectors (free-riders), and 2% mis-matchers. The distribution of types is very similar to the ones e.g. reported in Nosenzo and Tufano (2017); Miettinen et al. (2020), for slightly different parameters of the stage game.

[7] Note that given that types were determined endogenously within each session, it was not always possible to form all three types of groups. In particular, we had one session without a prosocial group and two sessions without a mixed group. This explains why the number of observations slightly differ across groups.

[8] See Figure A1 in Appendix A for a breakdown of cooperation rates by matching group.

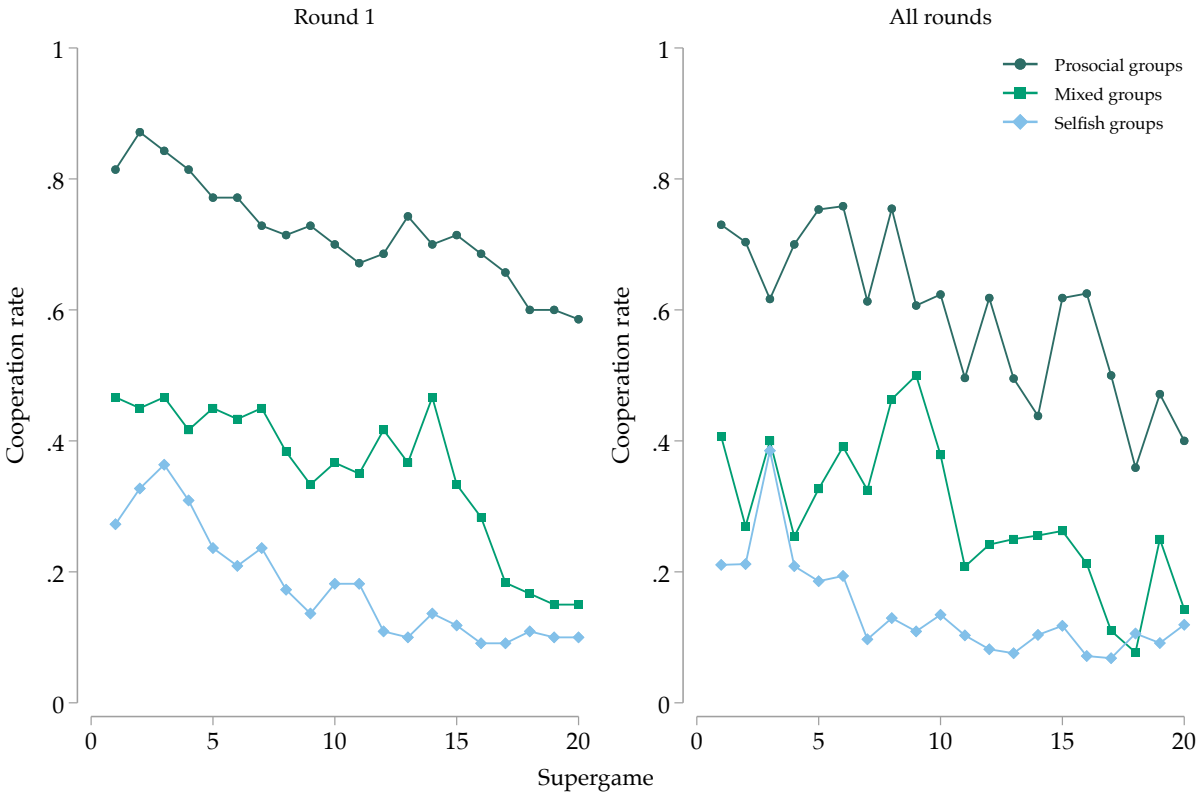

Figure 1: Evolution of cooperation by group type

ation throughout the entire game, in selfish groups we find cooperation rates to converge to very low levels towards the end of the game, which is in line with Hypothesis 1.[9] Averaged over all supergames, first round cooperation rates amount to 72% in prosocial groups, four times higher than in selfish groups (18%). Very similar differences are obtained when considering data from all rounds (prosocial groups: 58%, selfish groups: 14%) or when looking at cooperation rates in the last round of each supergame excluding those supergames that only lasted one round (see Figure A2 and Table A2 in Appendix A). This indicates that the differences in cooperation we observe do not only arise at the beginning of a supergame, but are maintained throughout longer repeated interactions.

Mixed groups start off with intermediate levels of cooperation that are higher than in

[9] Note that even in prosocial groups we observe a slight negative trend in cooperation. Given this time trend, one interesting question is whether the differences in cooperation rates would eventually disappear if the game is played long enough. To address this question, we follow the approach by Kartal and Müller (2018) who, based on a technique proposed by Noussair et al. (1995) and Barut et al. (2002), estimate asymptotes of cooperation rates. The results of this analysis reveal that cooperation rates in prosocial groups would not have converged to zero in the long run, but would have stayed significantly higher than in the other two types of groups (see Table A3 in Appendix A).

Table 3: Cooperation rates across supergames and group type

Cooperation rates 1st round	Supergame		
	1	20	1-20
Prosocial groups	0.81	0.59	0.72
Selfish groups	0.27	0.10	0.18
Mixed groups	0.47	0.15	0.35
H_0: Prosocial = Selfish	$p < 0.001$	$p < 0.001$	$p < 0.001$
H_0: Prosocial = Mixed	$p < 0.001$	$p = 0.005$	$p = 0.010$
H_0: Mixed = Selfish	$p = 0.007$	$p = 0.685$	$p = 0.144$
Cooperation rates all rounds			
Prosocial groups	0.73	0.40	0.58
Selfish groups	0.21	0.12	0.14
Mixed groups	0.41	0.14	0.29
H_0: Prosocial = Selfish	$p < 0.001$	$p = 0.016$	$p < 0.001$
H_0: Prosocial = Mixed	$p = 0.006$	$p = 0.073$	$p = 0.004$
H_0: Mixed = Selfish	$p < 0.001$	$p = 0.810$	$p = 0.208$

Notes: Differences between treatments are tested using probit regressions with standard errors clustered at the matching group level.

selfish groups but lower than in prosocial groups. As the game proceeds, however, cooperation rates decrease to similar levels than in selfish groups. Averaged over all supergames first (all) round cooperation rates amount to 35% (29%), which is only slightly higher than the levels observed in selfish groups, but much lower than in prosocial groups. This is in line with our Hypothesis 3.

To test the significance of these results, we use probit regressions with the decision to cooperate as the dependent variable, and dummy variables for the different group types as independent variables. In all regressions we cluster standard errors at the matching group level. The results are shown in Table 3, reporting the cooperation levels for the first, the last, and all supergames combined, along with the significance levels from each pairwise comparison. The upper panel shows cooperation levels using first round data, while the lower panel reports the results from all rounds.

The results in Table 3 confirm the visual impressions from Figure 1. In particular, they reveal that the difference between prosocial groups and the other two group types are not only large in size but also statistically significant right from the start of the game. The table further shows that these differences remain statistically significant and similar in size when considering only the last supergame or all supergames combined. This reveals that

experience or learning does not mitigate our observed differences. Finally, the table shows that while mixed groups cooperate at significantly higher levels than selfish groups at the beginning of the game, this difference becomes insignificant both when looking only at the last supergame or all supergames combined.

To further highlight the importance of group composition on cooperation, it is instructive to separately consider the behavior of prosocial and selfish types in mixed groups, and to compare it to the ones of their counterparts in segregated groups. To illustrate this, Figure 2 displays the evolution of cooperation in mixed groups, separately for both types. The left panel depicts first round cooperation rates and the right panel shows data from all rounds. The figure reveals that, despite having revealed very different attitudes towards cooperation in part 1 of the experiment, prosocial and selfish types cooperate to a remarkably similar degree when grouped together in mixed groups. Averaged over all supergames, first round cooperation rates amount to 40% for prosocial types and 31% for selfish types. Probit regressions reveal that this difference is not statistically significant ($p = 0.492$). A similar picture emerges when considering choices from all rounds. In this case, cooperation rates amount to 33% and 24%, respectively, which, again, is not statistically significant ($p = 0.449$).

When comparing these numbers to those observed in segregated groups, it becomes evident that while both types of players adjust their level of cooperation towards that of their counterpart, this adjustment is particularly strong for prosocial types. Specifically, compared to when interacting in segregated groups, prosocial types significantly decrease their cooperation level from 72% to 40% ($p = 0.068$) in mixed groups. For selfish types, in contrast, we only observe a slight but insignificant upward adjustment of cooperation rates, from 18% to 31% ($p = 0.225$). This asymmetric adjustment is consistent with our finding that about 90% of the prosocial types are conditional cooperators, i.e., they are only willing to cooperate if others do so, too.

In sum, our results reveal that social preferences do play a role in determining behavior in infinitely repeated games, but only if players interact in segregated groups. We summarize these findings in our first result:

> **Result 1:** *Grouping people with social preferences together in segregated groups has a strong positive effect on cooperation rates. When grouped with selfish types in mixed groups, however, cooperation rates are low and not significantly different from the ones observed in groups of purely self-interested players.*

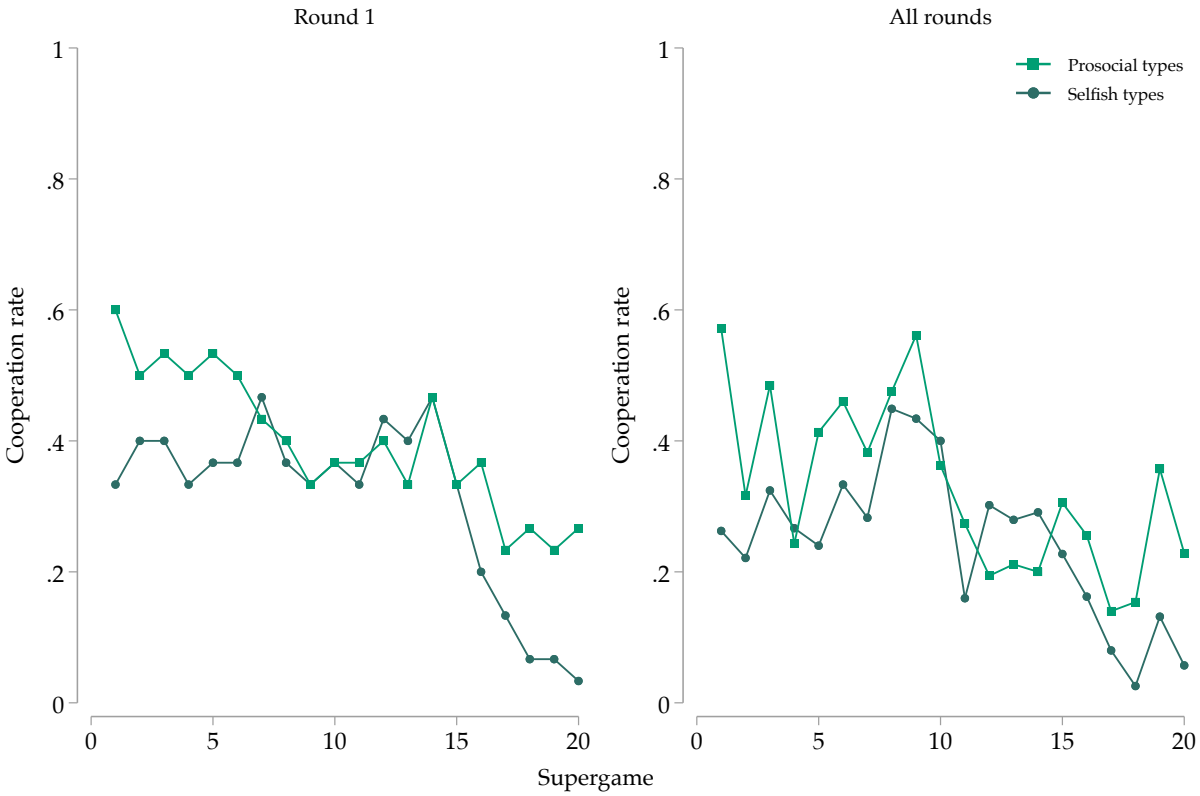

Figure 2: Cooperation in mixed groups by type.

To put our results into perspective, we can compare the achieved levels of cooperation in the different groups to those observed in previous studies. As shown in a recent survey by Dal Bó and Fréchette (2018), in the nine studies that have considered a situation with $\delta < \delta^{SPE}$, first round cooperation rates in the seventh supergame vary between 2% and 42%. The results from our mixed and selfish groups are roughly within that range, amounting to 45% and 24%, respectively. The levels observed in our prosocial groups, in contrast, are by far the highest ever reported in this type of context, amounting to 72%, more than three times higher than the median of 21% of these earlier studies.[10]

Yet, our results also show that while prosocial groups manage to achieve very high levels of cooperation, cooperation rates are below the social optimum of full cooperation. Recall that since almost all of our prosocial types are conditional cooperators, i.e., they are only willing to cooperate if others (are believed to) do so too, mutual cooperation and mutual

[10] To facilitate comparison of our results to those provided in Dal Bó and Fréchette (2018), we can normalize our game as displayed in Table 2 such that it can be described by the following two parameters: g, the gain from defection when the other player cooperates, and l, the loss from cooperation when the other player defects. In our case, we have $g = l = 2$.

defection are both equilibrium outcomes. As a consequence, depending on subjects' beliefs about others' behavior, prosocial groups might coordinate on different equilibria. Strategic uncertainty with regard to the partner's choice can explain why cooperation rates in prosocial groups do not reach 100%, and raises the question of what is the role for initial beliefs about others' willingness to cooperate for repeated game behavior in prosocial groups. We shed some light on this question by asking whether optimistic beliefs, as proxied by willingness to cooperate in the unconditional decision of part 1, predict cooperation in the repeated game. Interestingly, we find that beliefs play a limited role: optimistic beliefs predict cooperation in the first round of the first supergame (92% vs. 71%, $\chi^2(1) = 5.137$ $p = 0.023$), but the effect becomes small and statistically insignificant from the second supergame onwards (see Appendix Figure A3).

4.3 Repeated-game strategies

To better understand the observed differences in cooperation rates across the different group types, and to investigate to what extent the different cooperation types (as elicited in part 1 of the experiment) follow different strategies, in the following we estimate repeated-game strategies using the Strategy Frequency Estimation Method (SFEM) as proposed by Dal Bó and Fréchette (2011). Using maximum likelihood, this approach allows to estimate the prevalence of a certain set of predetermined strategies. Here we consider the five strategies that previous literature has identified to be the most relevant ones, accounting for a large majority of chosen strategies across a variety of settings (Dal Bó and Fréchette, 2019). The five strategies are: Always Defect (AD), Always Cooperate (AC), Grim trigger (GT), Tit-for-Tat (TFT), and suspicious Tit-for-Tat (STFT). These are standard strategies, except for STFT, which starts by defecting and, from then on, matches what the other player did in the previous round (see Appendix B for a detailed description of all strategies and estimation procedures).

Table 4 reports the estimates of the proportion for each strategy, separately for each group type. For mixed groups, we also estimate the frequency of strategies separately for prosocial and selfish types.[11] Table 4 reveals several interesting patterns. First, in line with

[11] One challenge of estimating strategies is that supergames which end after just one round do not allow to distinguish between cooperative strategies. As a robustness check, in the appendix we rerun our analysis using data from only those supergames with more than one round of interactions. We find that this slightly improves the efficiency of the estimation and reduces the trembling probability *Gamma*, but that this does not qualitatively affect our point estimates (see Table A4 in Appendix A). As an additional robustness check, we rerun our analysis by excluding the data from the first ten supergames, thus focusing only on data from

Table 4: Estimated strategy frequencies

	Prosocial groups	Mixed groups			Selfish groups
		All	Prosocial	Selfish	
Always defect (AD)	0.264***	0.679***	0.662***	0.700***	0.704***
	(0.093)	(0.141)	(0.162)	(0.151)	(0.075)
Always cooperate (AC)	0.091	0.033	0.035	0.031	0.003
	(0.071)	(0.031)	(0.039)	(0.023)	(0.013)
Grim trigger (GT)	0.199***	0.095	0.099	0.085	0.061
	(0.079)	(0.075)	(0.068)	(0.081)	(0.044)
Tit-for-tat (TFT)	0.395***	0.166**	0.205**	0.133	0.056
	(0.107)	(0.078)	(0.103)	(0.081)	(0.038)
Suspicious Tit-for-tat (STFT)	0.051	0.027	0.000	0.051	0.177***
	(0.038)	(0.041)	(0.022)	(0.065)	(0.074)
Gamma	0.483***	0.493***	0.465***	0.519***	0.436***
	(0.052)	(0.055)	(0.070)	(0.071)	(0.056)
Frequency of cooperative strategies	0.736	0.321	0.338	0.300	0.296
Observations	70	60	30	30	110

Notes: Estimates from maximum likelihood based on all rounds of all supergames. Bootstrap standard errors in parentheses. Statistical significance is assessed using Wald tests in which the null hypothesis is that a strategy frequency equals zero. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Gamma* is a positive coefficient that captures the amount of noise in the decision making. The frequency of cooperative strategies is equal to 1 minus the share AD.

previous findings for situations in which δ is not conducive to cooperation, we find that in mixed and selfish groups AD is the most common strategy, amounting to 68% and 70%, respectively (compare Dal Bó and Fréchette, 2011). The only other strategy for which the estimated frequency significantly differs from zero is TFT for mixed groups and STFT for selfish groups. A closer inspection of the strategies played in mixed groups reveals that the positive fraction of TFT is primarily driven by prosocial rather than selfish types. Yet, despite this difference, overall prosocial and selfish types display remarkably similar patterns of chosen strategies when matched in mixed groups.

These patterns change dramatically when considering prosocial groups. Here, a large majority of decisions (74%) are best described by cooperative strategies. Among the set of cooperative strategies, TFT is the most popular amounting to 40%, followed by GT with 20%. The unconditional cooperative strategy AC, in contrast, is chosen in only 9% of the cases. Hence, consistent with their choices from the first part of the experiment in which the large majority of prosocial types revealed that they are willing to cooperate only conditionally,

those supergames in which subjects already have gained experience in the game. The results, reported in Table A5 in Appendix A, are qualitatively very similar to the ones reported in the main text.

also in the repeated game they mainly choose conditionally cooperative strategies, with TFT being the modal one.

These results highlight that prosocial types adopt very different strategies depending on the type of group they are interacting in. In particular, while when matched among themselves the share of cooperative strategies amounts to almost three quarters, this number drops to one third when matched with selfish types in mixed groups. No such adjustment is observed for selfish types: irrespective of whether interacting in mixed or segregated groups, the estimated share of AD is 70%. We summarize these findings in our second result:

> **Result 2:** *Subjects classified as selfish choose predominantly non-cooperative strategies irrespective of the type of people they are matched with. Subjects classified as prosocial, in contrast, choose predominantly cooperative strategies, but only when matched among themselves.*

4.4 Discussion

So far, our results have provided strong evidence for the importance of social preferences and matching for the level of achievable cooperation in indefinitely repeated games. We have demonstrated that it requires both prosocial types and segregated groups in order for cooperation to thrive. This can explain why previous literature has not detected any pronounced effects of social preferences in repeated games, as in mixed groups cooperation ultimately breaks down.

In order to test our main hypothesis, we have relied on a situation in which the shadow of the future was not sufficiently long for purely self-interested players to have an incentive to cooperate. This had the major advantage that we could derive clean theoretical predictions for the different types of group.

A natural next question is whether social preferences play a similar role in environments that are instead favorable to cooperation. Previous literature has indicated that social preferences can help explain cooperation *only* when the material incentives of the game are *not* conducive to cooperation (Dreber et al., 2014). Moreover, while it has been shown that cooperation is higher when the game is conducive to cooperation, substantial variation remains, indicating that subjects do not necessarily coordinate on the Pareto-efficient equilibrium (Dal Bó and Fréchette, 2018). In the next section, we report the results from a follow-up experiment in which we test whether sorting people according to their prosocial attitudes also has an enhancing effect on cooperation when $\delta > \hat{\delta}^{SPE}$.

5 The role of social preferences when cooperation can be sustained even among self-interested players

Our new experiment was conducted with $n = 270$ participants. The design and the procedures were exactly the same as in our previous experiment, except that in part 2 of the experiment we set the continuation probability to $\delta = 0.8 > \hat{\delta}^{SPE} = 0.67$.[12] We thus have a setting in which cooperation can be sustained in equilibrium even among purely self-interested players.[13] Hence, without further assumptions, theory makes no clear prediction about potential differences in cooperation across the different group types as in each of them, both cooperation and defection constitute an equilibrium outcome. Yet, social preferences might serve as an equilibrium selection device, rendering this additional experiment interesting from an empirical point of view.

Based on subjects' choices in the first part of the experiment, we classify 43% as prosocial and 57% as selfish types. These numbers are very similar to the ones from our previous experiment. We had eight groups consisting of only prosocial types, eleven groups consisting of only selfish types, and eight mixed groups. In the latter, the average share of prosocial and selfish types was 44% and 56%, respectively.

Figure 3 displays the evolution of cooperation rates across the different types of groups using data from first rounds (left panel) and all rounds (right panel). The results reveal the power of social preferences in promoting cooperation even in this context. That is, we find cooperation to be considerably higher in prosocial groups compared to selfish groups. Averaged over all supergames, first (all) round cooperation rates amount to 85% (75%) in prosocial groups and 40% (26%) in selfish groups. As we show in Table 5, this difference is highly significant ($p < 0.001$). As before, we find cooperation rates in mixed groups to lie in between these two extremes. Differently from above, however, cooperation levels in mixed groups are now much closer to the ones in prosocial groups. In some cases (e.g. when considering data from all rounds), the difference between mixed and prosocial groups becomes statistically insignificant ($p = 0.187$). At the same time, cooperation levels in mixed groups are now significantly higher than in selfish groups, both when considering data from first rounds ($p = 0.013$) and from all rounds ($p < 0.001$). Figure 3 further reveals that

[12] This experiment was also pre-registered on the AsPredicted platform (#28887).

[13] Given the parameters of our game ($g = 2, l = 2$), cooperation is also risk-dominant (Blonski et al., 2011), with a basin of attraction (or $SizeBAD$) of 0.5 Dal Bó and Fréchette (2011). The latter index is defined as the maximum probability of the other player following the grim trigger strategy such that defection is optimal. Hence, in our case a purely self-interested player needs to believe that the other player will cooperate with a probability of at least 0.5 in order for him to also want to cooperate.

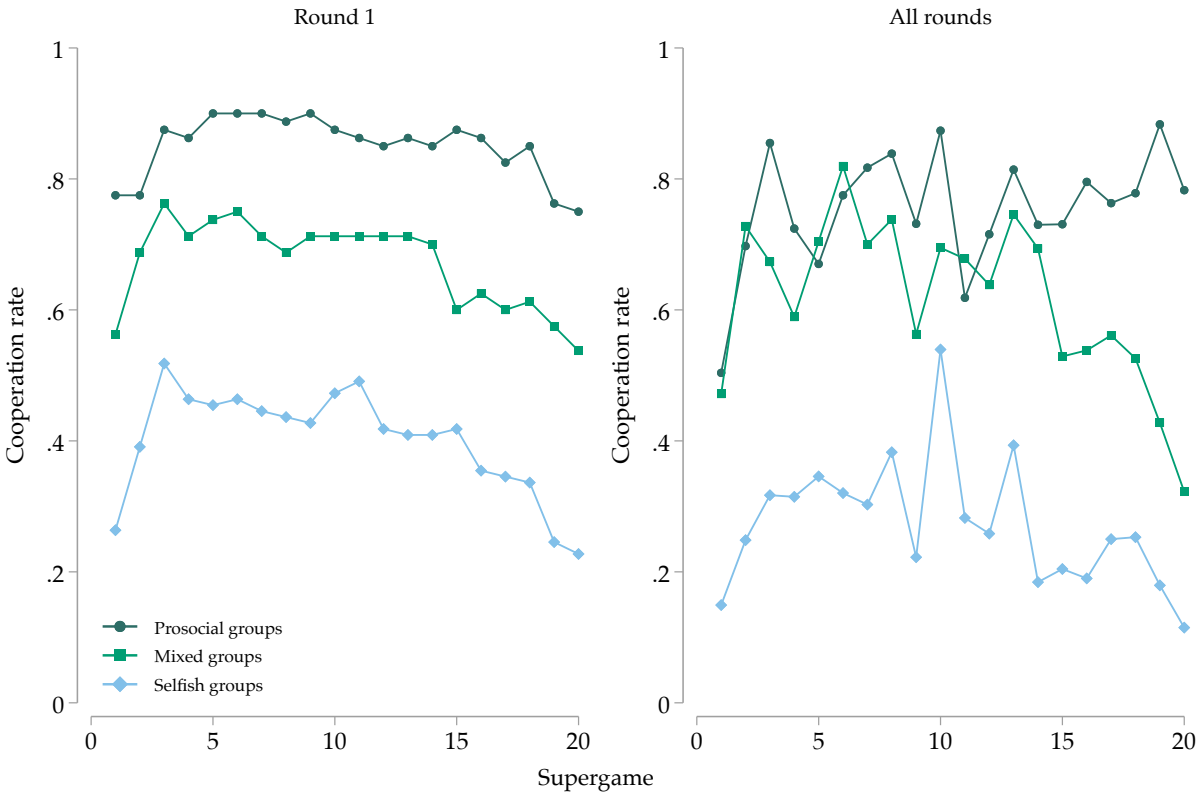

Figure 3: Evolution of cooperation for $\delta = 0.8$

the cooperation differences across groups are relatively stable across supergames, suggesting that, as in our first experiment, experience or learning effects do not mitigate our observed differences. We summarize these findings in our third result:

> **Result 3:** *Even when cooperation can be sustained in equilibrium among self-interested players, compared to groups of selfish players, grouping people with social preferences together has a strong positive effect on cooperation rates.*

To shed some further light on these results, analogously to the analysis above, we estimate repeated game strategies using the SFEM. The results from this estimation can be found in Table 6. It reveals that while in selfish groups the estimated frequency of cooperative strategies is only about 51%, this frequency increases to 79% in mixed groups and 90% in prosocial groups. The most common cooperative strategy across all groups is TFT, followed by GT. In prosocial and mixed groups, we further observe a significant share of AC, while in selfish groups we observe a significant share of STFT. The fact that selfish groups predominantly use non-cooperative or rather pessimistic strategies help explain why

Table 5: Cooperation rates across supergames and group type for $\delta = 0.8$

Cooperation rates 1st round	Supergame		
Group type	1	20	1-20
Prosocial groups	0.77	0.75	0.85
Selfish groups	0.26	0.23	0.40
Mixed groups	0.56	0.54	0.67
H_0: Prosocial = Selfish	$p < 0.001$	$p < 0.001$	$p < 0.001$
H_0: Prosocial = Mixed	$p < 0.006$	$p = 0.143$	$p = 0.079$
H_0: Mixed = Selfish	$p < 0.001$	$p = 0.021$	$p = 0.013$
Cooperation rates all rounds			
Prosocial groups	0.50	0.78	0.75
Selfish groups	0.15	0.11	0.26
Mixed groups	0.47	0.32	0.61
H_0: Prosocial = Selfish	$p = 0.004$	$p < 0.001$	$p < 0.001$
H_0: Prosocial = Mixed	$p = 0.796$	$p < 0.001$	$p = 0.187$
H_0: Mixed = Selfish	$p < 0.001$	$p = 0.008$	$p < 0.001$

Notes: Differences between treatments are tested using probit regressions with standard errors clustered at the matching group level.

they largely fail to coordinate on the efficient outcome of mutual cooperation. The high prevalence of cooperative strategies in mixed and prosocial groups, in contrast, can explain why they are very successful in achieving high levels of cooperation.

When comparing cooperation rates across our two experiments, we find that increasing δ from 0.6 to 0.8 has an overall positive effect on cooperation. This increase is particularly pronounced for mixed groups, which increase cooperation rates across all rounds by 32 percentage points (from 29% to 61%), compared to 17 and 12 percentage points in prosocial and selfish groups, respectively. This is further reflected by a shift in the repeated game strategies. Across all groups, we find that the estimated share of AD decreases under $\delta = 0.8$. In addition, we find that among the set cooperative strategies, the strategies played become more forgiving and optimistic. That is, we find that in all group types the combined relative share of AC and TFT increases, while the one for GT and STFT decreases.

In sum, our results reveal that even in a situation in which cooperation can be sustained as an equilibrium outcome among purely self-interested players, forming segregated groups of prosocial players has a strong positive effect on cooperation. This suggests that social preferences may serve as an important equilibrium selection device in repeated games.

Table 6: Estimated strategy frequencies for $\delta = 0.8$

	Prosocial groups	Mixed groups			Selfish groups
		All	Prosocial	Selfish	
Always defect (AD)	0.100**	0.215***	0.057	0.347***	0.491***
	(0.041)	(0.066)	(0.085)	(0.087)	(0.067)
Always cooperate (AC)	0.141***	0.213***	0.383***	0.090***	0.000
	(0.051)	(0.051)	(0.096)	(0.031)	(0.002)
Grim trigger (GT)	0.243***	0.179***	0.258***	0.111***	0.062***
	(0.067)	(0.048)	(0.061)	(0.042)	(0.021)
Tit-for-tat (TFT)	0.515***	0.322***	0.302***	0.325***	0.375***
	(0.072)	(0.060)	(0.055)	(0.100)	(0.068)
Suspicious Tit-for-tat (STFT)	0.000	0.071***	0.000	0.126***	0.072***
	(0.000)	(0.033)	(0.000)	(0.043)	(0.025)
Gamma	0.401***	0.389***	0.321***	0.444***	0.446***
	(0.033)	(0.019)	(0.029)	(0.029)	(0.028)
Frequency of cooperative strategies	0.900	0.785	0.943	0.653	0.509
Observations	80	80	35	45	110

Notes: Estimates from maximum likelihood based on all rounds of all supergames. Bootstrap standard errors in parentheses. Statistical significance is assessed using Wald tests in which the null hypothesis is that a strategy frequency equals zero. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Gamma* is a positive coefficient that captures the amount of noise in the decision making. The frequency of cooperative strategies is equal to 1 minus the share AD.

6 Validity of our type classification

Our findings have revealed strong differences in cooperation rates across prosocial and selfish groups, both when δ is conducive to cooperation and when it is not. We have argued that these differences are caused by differences in social preferences as elicited by the strategy method. Previous research has demonstrated that the strategy method is a behaviorally valid instrument to elicit subjects' attitudes (Fischbacher and Gächter, 2010; Brandts and Charness, 2011; Fischbacher et al., 2012; Gächter et al., 2017). In this section we use our post-experimental questionnaire to provide further evidence in support of our interpretation.

We start with the results from the norm following task. Kimbrough and Vostroknutov (2016, 2018) have shown that behavior in this task predicts social behavior in a variety of contexts, including dictator-game giving and second-mover behavior in a trust game. Our data reveals that behavior in this task is strongly correlated to subjects' revealed preferences for cooperation as elicited in part 1 of the experiment. This is highlighted by Figure 4, depicting the distribution of balls that were put in the prescribed blue urn, separately for subjects being classified as prosocial and selfish. The figure reveals pronounced differences

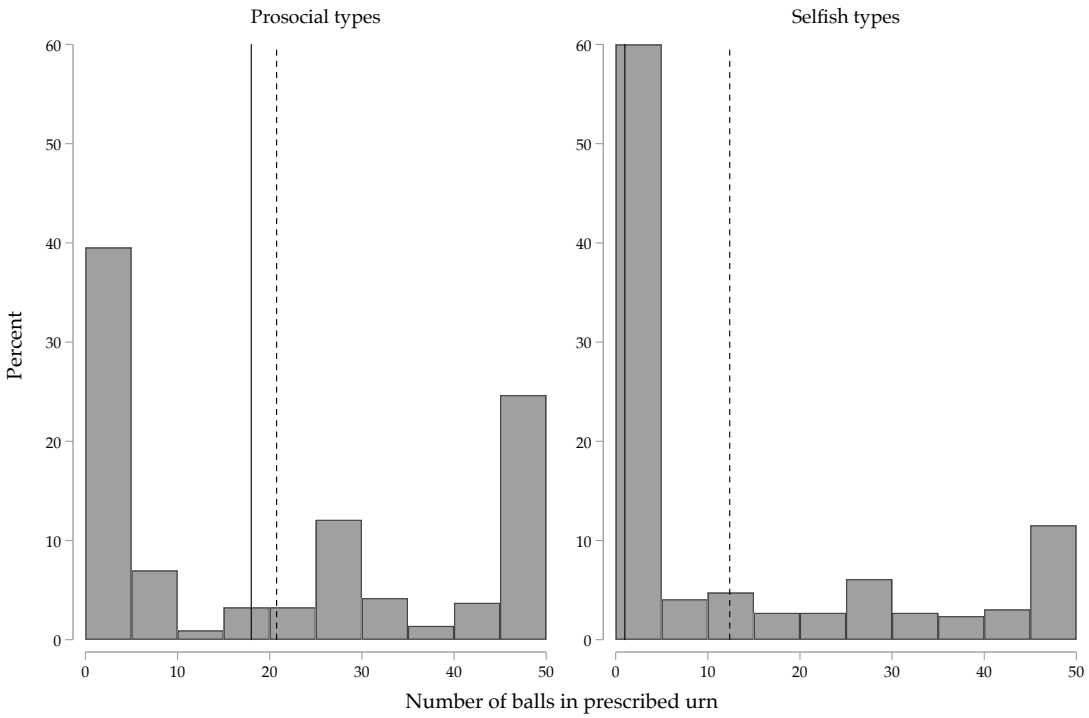

Figure 4: Distribution of the willingness to follow norms. Dashed lines display mean scores, solid lines display the median.

across the two samples, with prosocial types being significantly more likely to follow the rule of putting the balls in the blue urn than selfish types (Kolmogorov-Smirnov test, $p < 0.001$). The average (median) number of balls that were put in the prescribed urn amounts to 20.8 (18) for prosocial and 12.4 (1) for selfish types (t-test, $p < 0.001$). This result suggests that cooperation is indeed associated with normative behavior, as recently argued by Fehr and Schurtenberger (2018).

Next, we turn to the results from the big-five personality inventory (Schupp and Gerlitz, 2008). Previous studies have provided a link between personality traits and preferences towards cooperation (Volk et al., 2012) as well as strategic game-play (Proto et al., 2019). In line with this previous evidence, we find that subjects classified as prosocial score significantly higher on agreeableness (t-test, $p = 0.003$) and significantly lower on conscientiousness (t-test, $p = 0.010$) than selfish types (see Table A7 in Appendix A).

Finally, we ask to what extent our type classification is related to cognitive ability, as measured by subjects' performance in a 10-item Raven's progressive matrices test (Raven, 2000). This is interesting as previous literature has highlighted the importance of cognitive

skills for a variety of determinants of economic outcomes such as risk aversion, patience, and rationality (Frederick, 2005; Heckman et al., 2006; Borghans et al., 2008; Burks et al., 2009; Oechssler et al., 2009; Dohmen et al., 2010; Benjamin et al., 2013; Gill and Prowse, 2016). Moreover, intelligence has been shown to foster cooperation in indefinitely repeated interactions (Proto et al., 2019). Our results show no systematic relationship between our type classification and performance in the Raven's test: prosocial and selfish types both solve on average 4.22 tasks correctly (t-test, $p = 0.993$; compare also Figure A4 in Appendix A). These results indicate that differences in cognitive abilities are unlikely to be the explanation for the different cooperation rates we observe between prosocial and selfish groups.[14]

In sum, these results further strengthen our interpretation that differences in social preferences are the main driver of the observed differences in cooperation.

7 Conclusions

Social interactions predominantly take place under the shadow of the future. Previous literature on infinitely repeated games has highlighted the primary role of self-interested strategic considerations in explaining outcomes. Using indefinitely repeated prisoner's dilemma games, this paper demonstrates the importance of social preferences for achieving efficient cooperative outcomes. We show that high levels of cooperation can be sustained against the strategic incentives to defect when prosocial individuals interact in segregated groups and there is common knowledge of that. At the same time, we show that the power of social preferences in promoting cooperation is reduced when players with heterogeneous degrees of prosociality interact in mixed groups. These findings are important because they reveal novel insights on behavior in infinitely repeated games, and further provide an explanation for the inconclusive evidence on the role of social preferences in previous literature.

Our study also provides more general insights on the importance of a common level of prosociality within groups for the success of organizations and societies. There are several mechanisms that enable the formation of groups with homogeneous social preferences. First of all, as people have a preference to interact with others that are similar to them (McPherson

[14] In Appendix A, we further analyse the relative importance of norm-following, personality, and cognitive abilities for the likelihood of displaying a certain cooperative attitude. Using regression analysis with standardized coefficients, we find that norm following has the strongest positive impact, followed by agreeableness. Conscientiousness, in contrast, has the biggest negative impact on the likelihood of being a prosocial type. The results further reveal that gender and the general willingness to take risks are unrelated to subjects' willingness to reciprocate others' cooperation.

et al., 2001; Currarini et al., 2009), such groups may form rather naturally. Alternatively, such groups may be formed via the help of extrinsic mechanisms such as costly screening devices. For example, firms can use socially beneficial commitments to create incentives for individuals to self-sort into different groups (Brekke et al., 2011; Grimm and Mengel, 2009; Hauge et al., 2019). Finally, if the formation of segregated groups is not possible, e.g., if groups already exist, organizations may use other mechanisms to "educate" people to prosociality. For example, previous studies have suggested that promoting integration (Goette et al., 2006), shaping people's identity (Akerlof and Kranton, 2000, 2005) or investing in a socially-minded culture (Ashraf and Bandiera, 2017) have a positive effect on prosocial behavior. Studying the extent to which such mechanisms affect behavior in indefinitely repeated games is an interesting avenue for future research.

References

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *The Quarterly Journal of Economics*, 115, 715–753.

——— (2005): "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19, 9–32.

Aoyagi, M., V. Bhaskar, and G. R. Fréchette (2019): "The impact of monitoring in infinitely repeated games: Perfect, public, and private," *American Economic Journal: Microeconomics*, 11, 1–43.

Aoyagi, M. and G. Fréchette (2009): "Collusion as public monitoring becomes noisy: Experimental evidence," *Journal of Economic Theory*, 144, 1135–1165.

Arechar, A. A., A. Dreber, D. Fudenberg, and D. G. Rand (2017): ""I'm just a soul whose intentions are good": The role of communication in noisy repeated games," *Games and Economic Behavior*, 104, 726–743.

Ashraf, N. and O. Bandiera (2017): "Altruistic capital," *American Economic Review*, 107, 70–75.

Barut, Y., D. Kovenock, and C. N. Noussair (2002): "A comparison of multiple-unit all-pay and winner-pay auctions under incomplete information," *International Economic Review*, 43, 675–708.

Battigalli, P. and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review*, 97, 170–176.

Benjamin, D. J., S. A. Brown, and J. M. Shapiro (2013): "Who is 'behavioral'? Cognitive ability and anomalous preferences," *Journal of the European Economic Association*, 11, 1231–1255.

Bernheim, B. D. (1994): "A theory of conformity," *Journal of Political Economy*, 102, 841–877.

Bigoni, M., M. Casari, A. Skrzypacz, and G. Spagnolo (2015): "Time horizon and cooperation in continuous time," *Econometrica*, 83, 587–616.

Blonski, M., P. Ockenfels, and G. Spagnolo (2011): "Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence," *American Economic Journal: Microeconomics*, 3, 164–92.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90, 166–193.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008): "The economics and psychology of personality traits," *Journal of Human Resources*, 43, 972–1059.

Brandts, J. and G. Charness (2011): "The strategy versus the direct-response method: a first survey of experimental comparisons," *Experimental Economics*, 14, 375–398.

Brekke, K. A., K. E. Hauge, J. T. Lind, and K. Nyborg (2011): "Playing with the good guys. A public good game with endogenous group formation," *Journal of Public Economics*, 95, 1111–1118.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009): "Cognitive skills affect economic preferences, strategic behavior, and job attachment," *Proceedings of the National Academy of Sciences*, 106, 7745–7750.

Camera, G. and M. Casari (2009): "Cooperation among strangers under the shadow of the future," *American Economic Review*, 99, 979–1005.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *The Quarterly Journal of Economics*, 117, 817–869.

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

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree–An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Cooper, D. J. and J. Kagel (2016): "Other-regarding preferences," *The Handbook of Experimental Economics*, 2, 217.

Currarini, S., M. O. Jackson, and P. Pin (2009): "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77, 1003–1045.

Dal Bó, P. (2005): "Cooperation under the shadow of the future: experimental evidence from infinitely repeated games," *American Economic Review*, 95, 1591–1604.

Dal Bó, P. and G. R. Fréchette (2011): "The evolution of cooperation in infinitely repeated games: Experimental evidence," *American Economic Review*, 101, 411–29.

Dal Bó, P. and G. R. Fréchette (2018): "On the Determinants of Cooperation in Infinitely Repeated Games: A Survey," *Journal of Economic Literature*, 56, 60–114.

Dal Bó, P. and G. R. Fréchette (2019): "Strategy Choice in the Infinitely Repeated Prisoner's Dilemma," *American Economic Review*, 109, 3929–52.

Davis, D., A. Ivanov, and O. Korenok (2016): "Individual characteristics and behavior in repeated games: an experimental study," *Experimental Economics*, 19, 67–99.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010): "Are risk aversion and impatience related to cognitive ability?" *American Economic Review*, 100, 1238–60.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," *Journal of the European Economic Association*, 9, 522–550.

Dreber, A., D. Fudenberg, and D. G. Rand (2014): "Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics," *Journal of Economic Behavior & Organization*, 98, 41–55.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–298.

Engle-Warnick, J. and R. L. Slonim (2006): "Learning to trust in indefinitely repeated games," *Games and Economic Behavior*, 54, 95–114.

Falk, A. and U. Fischbacher (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fehr, E. and U. Fischbacher (2003): "The nature of human altruism," *Nature*, 425, 785.

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

——— (2006): "The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and new Theories," in *Handbook of the Economics of Giving, Altruism, and Reciprocity*, ed. by S. Kolm and J. M. Ythier, Elsevier, vol. 1, 615–691.

Fehr, E. and I. Schurtenberger (2018): "Normative foundations of human cooperation," *Nature Human Behaviour*, 2, 458.

Fischbacher, U. and S. Gächter (2010): "Social preferences, beliefs, and the dynamics of free riding in public goods experiments," *American Economic Review*, 100, 541–56.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71, 397–404.

Fischbacher, U., S. Gächter, and S. Quercia (2012): "The behavioral validity of the strategy method in public good experiments," *Journal of Economic Psychology*, 33, 897–913.

Fréchette, G. R. and S. Yuksel (2017): "Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination," *Experimental Economics*, 20, 279–308.

Frederick, S. (2005): "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 19, 25–42.

Fudenberg, D. and E. Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, 54, 533–554.

Fudenberg, D., D. G. Rand, and A. Dreber (2012): "Slow to anger and fast to forgive: Cooperation in an uncertain world," *American Economic Review*, 102, 720–49.

Gächter, S., F. Kölle, and S. Quercia (2017): "Reciprocity and the tragedies of maintaining and providing the commons," *Nature Human Behaviour*, 1, 650–656.

Gächter, S. and C. Thöni (2005): "Social learning and voluntary cooperation among like-minded people," *Journal of the European Economic Association*, 3, 303–314.

Galizzi, M. M. and D. Navarro-Martínez (2019): "On the external validity of social preference games: a systematic lab-field study," *Management Science*, 65, 976–1002.

Ghidoni, R. and S. Suetens (2019): "Empirical evidence on repeated sequential games," Tech. rep., CEPR Discussion Papers.

Gill, D. and V. Prowse (2016): "Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis," *Journal of Political Economy*, 124, 1619–1676.

Gintis, H. (2000): "Strong reciprocity and human sociality," *Journal of Theoretical Biology*, 206, 169–179.

Goette, L., D. Huffman, and S. Meier (2006): "The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups," *American Economic Review*, 96, 212–216.

Götte, L. and E. Tripodi (2018): "Social Influence in Prosocial Behavior: Evidence from a Large-Scale Experiment," Tech. rep., CEPR Discussion Papers.

Greiner, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Grimm, V. and F. Mengel (2009): "Cooperation in viscous populations—Experimental evidence," *Games and Economic Behavior*, 66, 202–220.

Hauge, K. E., K. A. Brekke, K. Nyborg, and J. T. Lind (2019): "Sustaining cooperation through self-sorting: The good, the bad, and the conditional," *Proceedings of the National Academy of Sciences*, 116, 5299–5304.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor economics*, 24, 411–482.

Kartal, M. and W. Müller (2018): "A new approach to the analysis of cooperation under the shadow of the future: Theory and experimental evidence," *Available at SSRN 3222964*.

Kimbrough, E. O. and A. Vostroknutov (2016): "Norms make preferences social," *Journal of the European Economic Association*, 14, 608–638.

——— (2018): "A portable method of eliciting respect for social norms," *Economics Letters*, 168, 147–150.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of Economic Dynamics*, 1, 593–622.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001): "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 27, 415–444.

Mengel, F. (2017): "Risk and Temptation: A Meta-study on Prisoner's Dilemma Games," *The Economic Journal*, 128, 3182–3209.

Miettinen, T., M. Kosfeld, E. Fehr, and J. Weibull (2020): "Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions," *Journal of Economic Behavior & Organization*, 173, 1–25.

Nosenzo, D. and F. Tufano (2017): "The effect of voluntary participation on cooperation," *Journal of Economic Behavior & Organization*, 142, 307–319.

Noussair, C. N., C. R. Plott, and R. G. Riezman (1995): "An Experimental Investigation of the Patterns of International Trade," *The American Economic Review*, 462–491.

Oechssler, J., A. Roider, and P. W. Schmitz (2009): "Cognitive abilities and behavioral biases," *Journal of Economic Behavior & Organization*, 72, 147–152.

Palfrey, T. R. and H. Rosenthal (1994): "Repeated play, cooperation and coordination: An experimental study," *The Review of Economic Studies*, 61, 545–565.

Proto, E., A. Rustichini, and A. Sofianos (2019): "Intelligence, Personality, and Gains from Cooperation in Repeated Interactions," *Journal of Political Economy*, 127, 000–000.

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

Raven, J. (2000): "The Raven's progressive matrices: change and stability over culture and time," *Cognitive Psychology*, 41, 1–48.

Reuben, E. and S. Suetens (2012): "Revisiting strategic versus non-strategic cooperation," *Experimental Economics*, 15, 24–43.

Romero, J. and Y. Rosokha (2018): "Constructing strategies in the indefinitely repeated prisoner's dilemma game," *European Economic Review*, 104, 185–219.

Roth, A. E. and J. K. Murnighan (1978): "Equilibrium behavior and repeated play of the prisoner's dilemma," *Journal of Mathematical Psychology*, 17, 189–198.

Schupp, J. and J.-Y. Gerlitz (2008): "BFI-S: big five inventory-SOEP," in *Zusammenstellung sozialwissenschaftlicher items und skalen. ZIS Version*, vol. 12, 7.

Selten, R. (1967): "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes," in *Beiträge zur experimentellen Wirtschaftsforschung*, Sauermann, H., 136–168.

Sobel, J. (2005): "Interdependent preferences and reciprocity," *Journal of Economic Literature*, 43, 392–436.

Volk, S., C. Thöni, and W. Ruigrok (2012): "Temporal stability and psychological foundations of cooperation preferences," *Journal of Economic Behavior & Organization*, 81, 664–676.

Appendix for Online Publication

A Additional Figures and Tables

Table A1: Part 1 behavior and types

	Conditional decision on other's defection, cooperation			
	Defect, Defect (*free riders*)	Cooperate, Defect (*mis-matchers*)	Defect, Cooperate (*conditional cooperators*)	Cooperate, Cooperate (*unconditional cooperators*)
Unconditional decision				
Defect	43.33	2.50	17.92	1.25
Cooperate	12.50	0.0	18.75	4.75
Total	55.83	2.50	36.67	5.00

Notes: The table reports the fraction of respondents (in percentage) for each possible combination of choices in part 1 of the experiment.

Table A2: Last round cooperation rates across supergames and group type ($\delta = 0.6$)

	Supergame		
	1	20	1-20
Prosocial groups	0.68	0.45	0.50
Selfish groups	0.30	0.10	0.19
Mixed groups	0.15	0.10	0.09
H_0: Prosocial = Selfish	$p < 0.001$	$p = 0.006$	$p < 0.001$
H_0: Prosocial = Mixed	$p = 0.003$	$p = 0.003$	$p = 0.026$
H_0: Mixed = Selfish	$p = 0.002$	$p = 1.000$	$p = 0.314$

Notes: Differences between treatments are tested using probit regressions with standard errors clustered at the matching group level.

Table A3: Estimated Asymptotes ($\delta = 0.6$)

Cooperation Rate, First Round		
Prosocial groups	Mixed groups	Selfish groups
0.678^{***}	0.311^{*}	0.134^{**}
(0.111)	(0.133)	(0.059)
$\beta_{Coop} = \beta_{Mixed}$	$\beta_{Mixed} = \beta_{Def}$	$\beta_{Coop} = \beta_{Def}$
$p = 0.131$	$p = 0.039$	$p < 0.001$
Cooperation Rate, All Rounds		
Prosocial groups	Mixed groups	Selfish groups
0.602^{***}	0.273^{*}	0.121^{**}
(0.112)	(0.135)	(0.047)
$\beta_{Coop} = \beta_{Mixed}$	$\beta_{Mixed} = \beta_{Def}$	$\beta_{Coop} = \beta_{Def}$
$p = 0.163$	$p = 0.023$	$p < 0.001$

Notes: Standard errors in parentheses are clustered at the matching group level. Reported p-values follow a two-sided z-test.

Table A4: Estimated strategy frequencies excluding supergames with only one round ($\delta = 0.6$)

	Prosocial groups	Mixed groups			Selfish groups
		All	Prosocial	Selfish	
Always defect (AD)	0.268***	0.662***	0.666***	0.664***	0.717***
	(0.102)	(0.133)	(0.154)	(0.115)	(0.068)
Always cooperate (AC)	0.110**	0.034	0.035	0.032	0.003
	(0.048)	(0.038)	(0.024)	(0.029)	(0.021)
Grim trigger (GT)	0.193**	0.106*	0.094**	0.109**	0.041
	(0.082)	(0.054)	(0.043)	(0.055)	(0.041)
Tit-for-tat (TFT)	0.408***	0.164**	0.205	0.133*	0.067**
	(0.082)	(0.076)	(0.137)	(0.078)	(0.034)
Suspicious Tit-for-tat (STFT)	0.021	0.034	0.000	0.063	0.172**
	(0.022)	(0.033)	(0.038)	(0.064)	(0.068)
Gamma	0.482***	0.449***	0.438***	0.457***	0.412***
	(0.045)	(0.048)	(0.078)	(0.070)	(0.050)
Frequency of cooperative strategies	0.732	0.338	0.334	0.336	0.283
Observations	70	60	30	30	110

Notes: Estimates from maximum likelihood based on all rounds of supergames with more than one round. Bootstrap standard errors in parentheses. Statistical significance is assessed using Wald tests in which the null hypothesis is that a strategy frequency equals zero. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Gamma* is a positive coefficient that captures the amount of noise in the decision making. The frequency of cooperative strategies is equal to 1 minus the share AD.

Table A5: Estimated strategy frequencies excluding the first 10 supergames ($\delta = 0.6$)

	Prosocial groups	Mixed groups			Selfish groups
		All	Prosocial	Selfish	
Always defect (AD)	0.306***	0.777***	0.729***	0.800***	0.674***
	(0.096)	(0.150)	(0.171)	(0.140)	(0.159)
Always cooperate (AC)	0.062	0.052	0.000	0.075	0.000
	(0.051)	(0.043)	(0.000)	(0.049)	(0.007)
Grim trigger (GT)	0.206**	0.093	0.064	0.105	0.079
	(0.081)	(0.078)	(0.077)	(0.077)	(0.060)
Tit-for-tat (TFT)	0.396***	0.079	0.184*	0.020	0.000
	(0.115)	(0.071)	(0.111)	(0.053)	(0.005)
Suspicious Tit-for-tat (STFT)	0.031	0.000	0.023	0.000	0.246
	(0.043)	(0.039)	(0.079)	(0.000)	(0.189)
Gamma	0.436***	0.402***	0.354***	0.441***	0.349***
	(0.058)	(0.052)	(0.070)	(0.065)	(0.058)
Frequency of cooperative strategies	0.694	0.223	0.271	0.200	0.326
Observations	70	60	30	30	110

Notes: Estimates from maximum likelihood based on all rounds of the last ten supergames (supergames 11 - 20). Bootstrap standard errors in parentheses. Statistical significance is assessed using Wald tests in which the null hypothesis is that a strategy frequency equals zero. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Gamma* is a positive coefficient that captures the amount of noise in the decision making. The frequency of cooperative strategies is equal to 1 minus the share AD.

Table A6: Estimated strategy frequencies excluding the first 10 supergames ($\delta = 0.8$)

	Prosocial groups	Mixed groups			Selfish groups
		All	Prosocial	Selfish	
Always defect (AD)	0.125***	0.229***	0.057	0.378***	0.497***
	(0.044)	(0.073)	(0.085)	(0.096)	(0.066)
Always cooperate (AC)	0.156***	0.227***	0.389***	0.103***	0.027
	(0.059)	(0.058)	(0.111)	(0.040)	(0.020)
Grim trigger (GT)	0.228***	0.191***	0.230***	0.146***	0.085**
	(0.066)	(0.057)	(0.068)	(0.054)	(0.040)
Tit-for-tat (TFT)	0.480***	0.241***	0.324***	0.175	0.276***
	(0.073)	(0.058)	(0.067)	(0.118)	(0.064)
Suspicious Tit-for-tat (STFT)	0.011	0.112*	0.000	0.199***	0.115***
	(0.011)	(0.057)	(0.000)	(0.069)	(0.033)
Gamma	0.344***	0.360***	0.312***	0.400***	0.385***
	(0.032)	(0.030)	(0.048)	(0.045)	(0.028)
Frequency of cooperative strategies	0.875	0.771	0.943	0.622	0.503
Observations	80	80	35	45	110

Notes: Bootstrap standard errors in parentheses. Statistical significance is assessed using Wald tests in which the null hypothesis is that a strategy frequency equals zero. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Gamma* is a positive coefficient that captures the amount of noise in the decision making. The frequency of cooperative strategies is given by the sum of frequencies except for AD.

Table A7: Predictors of decision to cooperate conditional on partner's cooperation

	(1)
Female	0.064
	(0.046)
Norm following	0.181***
	(0.001)
Raven test score	0.035
	(0.011)
General risk attitude	-0.035
	(0.010)
Agreeableness	0.115***
	(0.020)
Conscientiousness	-0.146***
	(0.022)
Extraversion	0.043
	(0.018)
Neuroticism	0.094**
	(0.017)
Openness	0.055
	(0.018)
Observations	510

Notes: OLS regression with standardized (beta) coefficients. Standard errors in parentheses are clustered at the individual level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

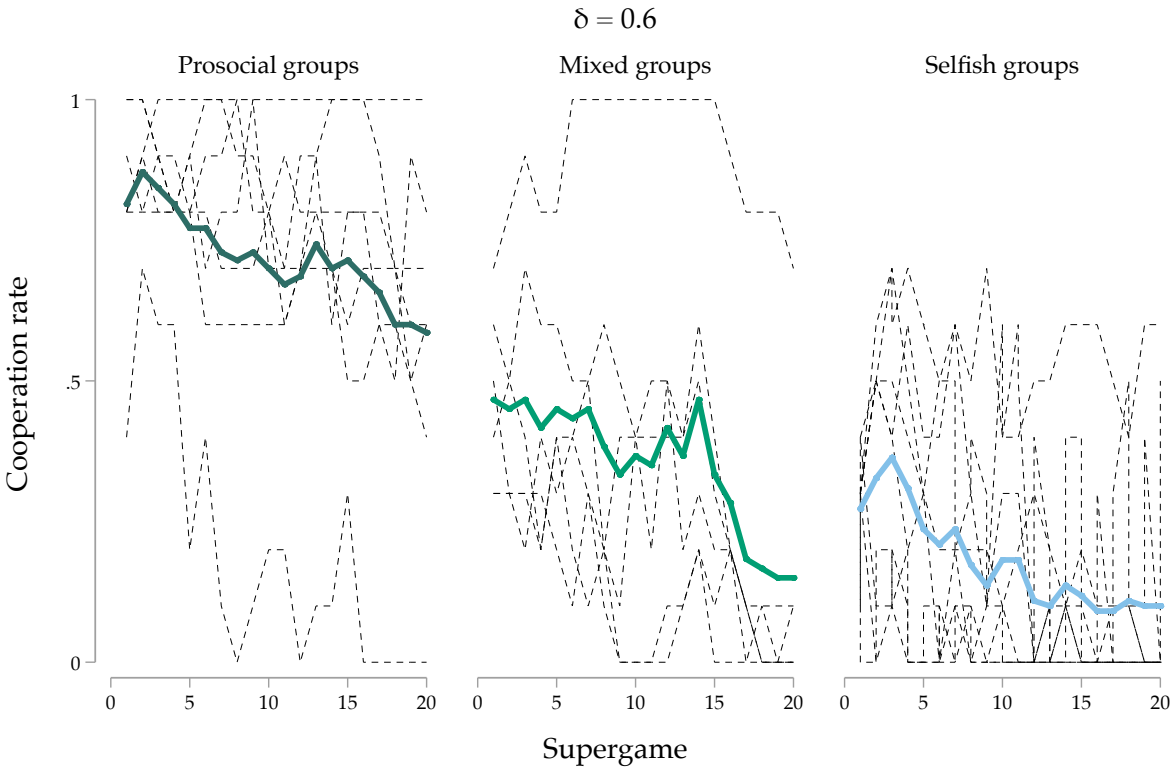

Figure A1: First round cooperation by matching groups (thick lines display overall averages)

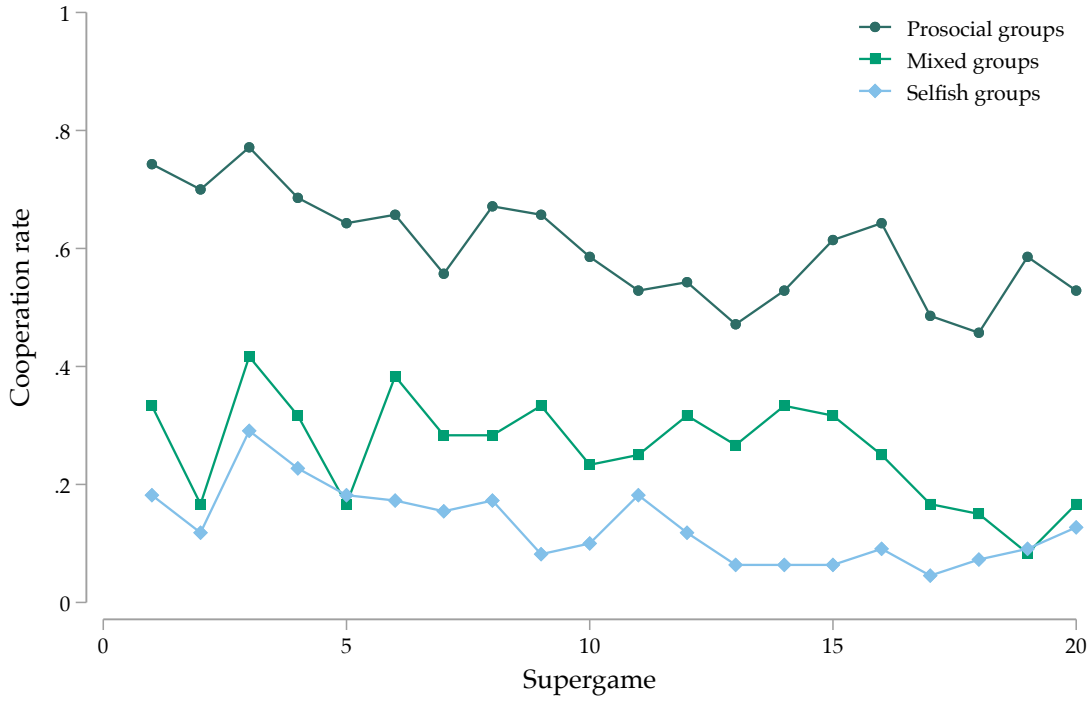

Figure A2: Evolution of last round cooperation by group type

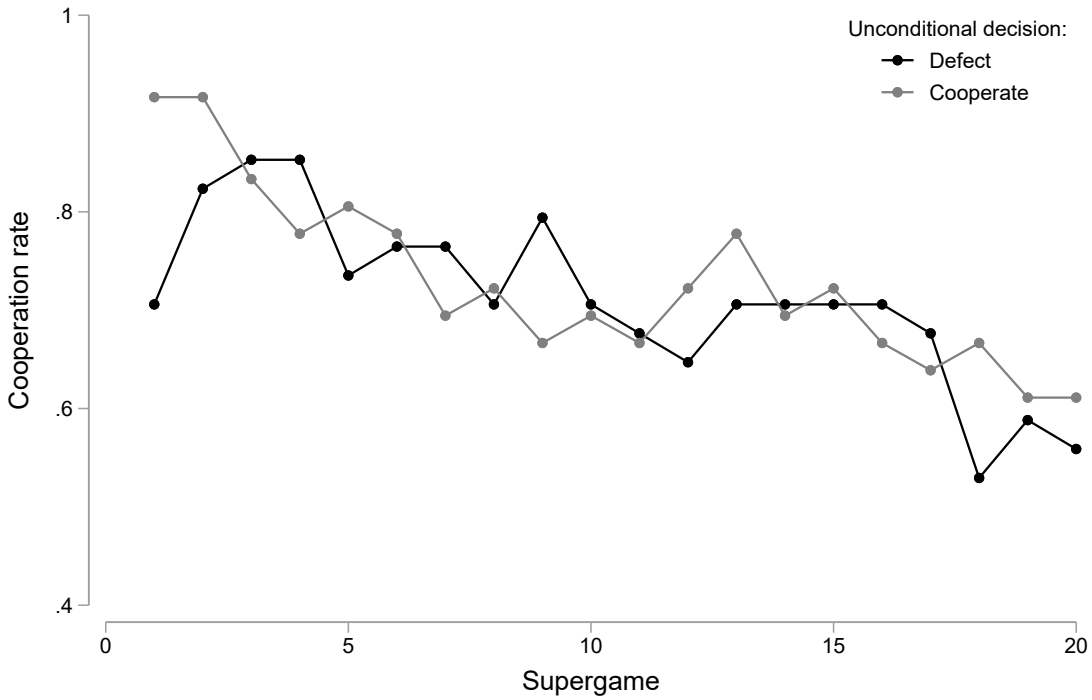

Figure A3: First round cooperation rates in prosocial groups by the unconditional cooperation decision in part 1

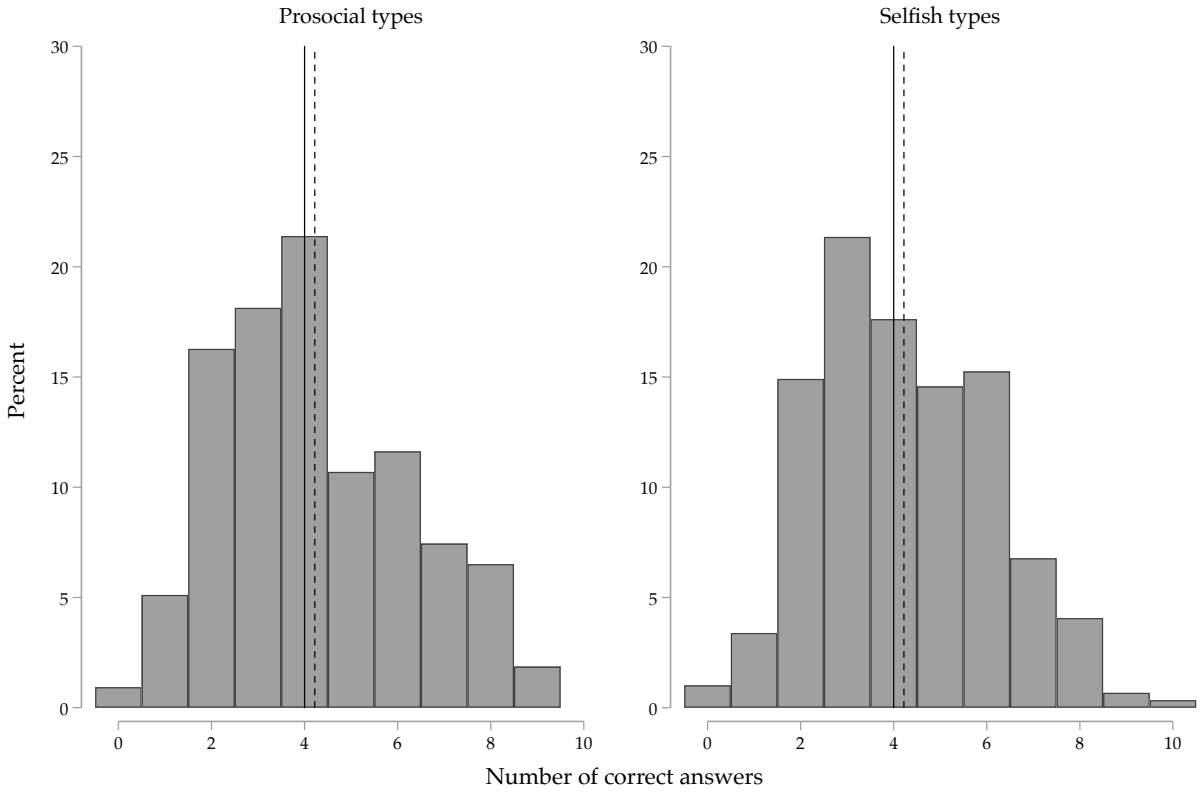

Figure A4: Distribution of number of correct answers in the IQ task. Dashed lines display the mean score, solid lines display the median.

B Stategy Frequency Estimation

To estimate the frequency of strategies chosen by players we follow the methodology proposed by Dal Bó and Fréchette (2011). We focus on memory-one strategies for their wide prevalence in studies that elicit them directly (see Romero and Rosokha, 2018; Dal Bó and Fréchette, 2019). We assume that individuals choose one strategy s^k, from a pool of five well established strategies (Dal Bó and Fréchette, 2019), and estimate the frequency that each of these strategies is played using maximum likelihood. The strategies are the following:

Strategy	Action
Always defect	$a_{i,m,r} = 0$
Always cooperate	$a_{i,m,r} = 1$
Grim	$a_{i,m,1} = 1, \quad a_{i,m,r>1} = 0 \quad$ if $\quad a_{j,m,r} = 0$ for some $r' < r$
Tit-for-tat	$a_{i,m,1} = 1, \quad a_{i,m,r>1} = a_{j,m,r-1}$
Suspicious Tit-for-tat	$a_{i,m,1} = 0, \quad a_{i,m,r>1} = a_{j,m,r-1}$

At each round $r \in R$, of each supergame $g \in G = \{1, 2, ..., 20\}$, an individual i chooses action $a_{i,g,r}$ that is either 1 (cooperate) or 0 (defect). An action is determined by a fully specified plan of actions–a strategy– that we allow to be followed by the player with some error $\varepsilon_{i,g,r}$. In turn, an action corresponds to

$$a_{i,g,r} = 1\{\tilde{a}_{i,g,r}(s^k) + \gamma\varepsilon_{i,g,r} \geq 0\}$$

where $\tilde{a}_{i,g,r}(s^k)$ is the latent choice implied by strategy s^k given the history of interactions in supergame g prior to round r.[15] Conditional on the strategy, actions across rounds and super-games are assumed to be independent, so that for an extreme value distributed error term one can write the likelihood (that player i cooperates) under strategy s^k in logistic form

$$p_i(s^k) = \prod^{G}\prod^{R}\left(\frac{1}{1+e^{-\tilde{a}_{i,g,r}(s^k)/\gamma}}\right)^{a_{i,g,r}}\left(\frac{1}{1+e^{\tilde{a}_{i,g,r}(s^k)/\gamma}}\right)^{1-a_{i,g,r}}.$$

For any i, the likelihood of cooperating under any of the strategies in the strategy consideration set K is $\sum^{K}\theta^k p_i(s^k)$. We use numerical methods to estimate the frequencies of each strategy θ^k and the variance γ of the error term $\varepsilon_{i,g,r}$, among individuals in a given

[15] For estimation convenience $\tilde{a}_{i,g,r}(s^k)$ is coded as 1 if the action prescribed by strategy s^k is cooperate, and -1 if the action prescribed by the strategy is defect.

sample, from the following log-likelihood function

$$l = \sum^{I} ln \left(\sum^{K} \theta^k p_i(s^k) \right).$$

C Experimental Instructions (Translated from German)

C.1 First Screen

General Information

Welcome and thank you for your participation in this experiment. For your participation and punctual arrival you receive 4€. You can earn an additional amount of money in this experiment. The exact amount you will receive depends on your decisions and the decisions of the other participants. It is therefore very important that you read the following instructions carefully.

General Rules

The results of this experiment will be used for a research project. It is therefore important that all participants follow certain rules of conduct. During the experiment, you are not allowed to communicate with other participants or any person outside the laboratory. For this reason, all mobile phones have to be switched off. If you have questions regarding the instructions or the study, please raise your hand – we will privately answer your question at your place. Disregarding this rule leads to the exclusion from this experiment and from all payments.

Anonymity

All decisions are made anonymously, i.e., no other participant learns about the identity of a participant who made a certain decision. Also, the payment is made anonymously, i.e., no participant learns about the payment of the other participants.

Course of the experiment

The experiment consists of two parts, which we will refer to as Part 1 and Part 2. At the end of the experiment, we will randomly select one of the two parts. Both parts have an equal probability of being selected. The selected part will then determine your earnings. Because you do not yet know which of the two parts will be relevant for your earnings, the best strategy, you should think about each decision carefully, since all of them can influence your earnings.

C.2 Part 1 Instructions

The decision situation

At first you will be informed about the general decision situation. Following that you will receive your task. At the beginning of Part 1, you will form a pair with another participant. Both you and the other participant can decide between two options, which we will call A and B. In the table below, you see the point earnings for you and the other participant depending on your and their choices.

		Other	
		A	**B**
You	**A**	You: 15€, Other: 15€	You: 0€, Other: 25€
	B	You: 25€, Other: 0€	You: 10€, Other: 10€

This means, that if:

- You choose A and the other participant chooses A, you each earn 15€.
- You choose A and the other participant chooses B, you earn 0€ and the other participant earns 25€.
- You choose B and the other participant chooses A, you earn 25€ and the other participant earns 0€.
- You choose B and the other participant chooses B, you each earn 10€.

Your task

The game is based on the decision situation described above. You and the other participant have to make two types of decisions, which we will refer to as the "unconditional decision" and the "conditional decision".

- In the unconditional decision you simply decide whether you choose A or B.
- In the conditional decision you can make your decision dependent on what the other participant in your group chose in their unconditional decision. That is, you can decide
  - whether you want to choose A or B in case the other participant chose A.

– whether you want to choose A or B in case the other participant chose B.

Once both you and the other participant have made both type of decisions, the computer program will randomly determine (with equal probability) which decision will be relevant for your earnings. In particular, for one participant in the pair the unconditional decision will be relevant to determine earnings, while for the other participant in the pair the conditional decision will be relevant to determine earnings. Which of the two conditional decisions is relevant then depends on the unconditional decision of the other participant. The following example makes this clear.

Example: There are two players, player 1 and player 2, and that the random mechanism determines that the unconditional decision is relevant for player 1 and that the conditional decision is relevant for player 2. Then, if player 1 chose option A in the unconditional decision, the relevant decision of player 2 will be determined by checking their conditional decision in case the other participant chose A. If in that case player 2 chose option B, he will earn 25€ and player 1 will earn 0€. If, instead, in that case player 2 chose option A, then both players will earn 15€.

—Earnings If Part 1 is selected at the end of the experiment to determine your earnings, you will receive that amount in cash.

After you make your choices in Part 1 you will proceed to Part 2. You will learn about others' decisions and your earnings from Part 1 only at the very end of the experiment, after Part 2 has ended. Note that your choices in Part 1 will determine with whom you are going to interact in Part 2. Please note, that in Part 2 you may either interact with participants who made the same Part 1 choices as you, or you will interact with participants who made different Part 1 choices. You will receive further information at the beginning of Part 2.

Control questions Part 1

We now ask you to answer a few questions, to make sure that all participants understand the instructions entirely.

Suppose you and the other participant chose the following decisions:

You:

- Unconditional decision: A
- Conditional decision if the other participants chooses A: A
- Conditional decision if the other participants chooses B: B

Other participant:

- Unconditional decision: A
- Conditional decision if the other participants chooses A: B
- Conditional decision if the other participants chooses B: B

The mechanism determines, that the unconditional decision is relevant for you and the conditional decision is relevant for the other participant.

1. What are your earnings?
2. What are the earnings of the other player?
   Now assume, that the mechanism determines, that the conditional decision is relevant for you and the unconditional decision is relevant for the other participant.
3. What are your earnings?
4. What are the earnings of the other player?

C.3 Part 2 Instructions

The decision situation

In the beginning of Part 2 you will again be paired with another participant. Both you and the other participant can choose between two options, A and B. Below, you find the table describing the earnings (in €) for you and the other participant depending on your and their choices.

		Other	
		A	**B**
You	**A**	You: 15€, Other: 15€	You: 0€, Other: 25€
	B	You: 25€, Other: 0€	You: 10€, Other: 10€

Your task

As you might have noticed, the table is the same as in Part 1. However, in contrast to Part 1, in Part 2 you will be asked to make decisions in several rounds. In each round you will only

have to make one decision between A and B. This is similar to the unconditional decision from Part 1. There is no conditional decision.

The timeline of Part 2 is as follows:

- First, you will be randomly paired with another participant to play the game above for a sequence of rounds. You will play with this same participant for the entire match.
- The length of the sequence is randomly determined. After each round, there is a 60% chance that the match will continue for at least one more round. So, for instance, if you are in round 2, the chance that there will be a third round is 60% and if you are in round 9, the chance that there will be another round is also 60%.
- Once a sequence ends, you will be randomly paired with another participant to play another sequence. In this new sequence, you will again play the same game for multiple rounds. After that, a new sequence with another randomly determined participant will be formed. There will be a total of 20 sequences

Important: in each sequence you are paired with a participant randomly drawn from a group of 9 people. The group of 10 people (including yourself) has been determined according to one of the conditional decisions of Part 1. In particular, all participants in your group (including you) choose to play A in case the other player chooses A.[16] all participants in your group (including you) choose to play B in case the other player chooses A.[17] some participants in your group choose A in case the other player choose A and some choose B in case the other player chooses A.[18]

Earnings

If Part 2 is chosen to determine your earnings, they are calculated as follows. One of the 20 sequences will be randomly selected. Your result in the last round of the chosen sequence will determine your earnings (in €). You will get to know the chosen sequence and your exact earnings at the very end of the experiment.

Control questions Part 2

We now ask you to answer a few questions, to make sure that all participants understand the instructions entirely.

[16] Displayed to subjects assigned to group where all members have decided to respond to cooperation with cooperation in the conditional decision of Part 1 of the experiment.

[17] Displayed to subjects assigned to group where all members have decided to respond to cooperation with defection in the conditional decision of Part 1 of the experiment.

[18] Displayed to subjects assigned to the mixed group.

1. How high is the probability after each round in a sequence that another round will be played (in %)?

2. Will you always interact with the same participant in a sequence?

3. Will you always interact with the same participants across sequences?

4. In Part 2 you exclusively play with participants who chose the following in Part 1 of the experiment:

   - Option A if the other participant chose option A
   - Option B if the other participant chose option A
   - Option A or option B if the other participant chose option A

5. What is the total number of sequences? Assume, that sequence 13 from Part 2 of the experiment was randomly chosen to determine youre earnings. Furthermore, assume that the sequence consisted of five rounds and the decisions were as follwing: You: A, A, A, B, B and the other participant: A, A, B, B, A

6. What are your earnings in this situation?

7. What are the earnings of the other participant in this situation?

D Norm Following Task

The norm following task was introduced by Kimbrough and Vostroknutov (2016) to investigate the possibility that prosocial behavior can be explained by an intrinsic desire to follow norms, and validate the method eliciting both the perception of norms and behavior for standard economic games: the public goods, trust, dictator, and ultimatum games.

The first proposed version of the norm following task leverages norms from outside the laboratory, such as pedestrian crossing with green/red light, to elicit an individual continuous measure of willingness to forgo monetary benefits to respect the rule. This task is charged of local perceptions of norms that may confound the measure of individual willingness to follow norms. We use a later version of the norm following task Kimbrough and Vostroknutov (2018) that is designed precisely to overcome this issue, by providing subjects with an explicit rule to follow in an abstract environment.

D.1 Experimental Instructions (translated from German)

In the following task you need to decide how to split 50 balls into two containers. Your task is to place each ball, one after the other, in one of the two containers: in the **blue** or the yellow container.

The balls will appear on your screen, and you can distribute each ball by clicking and dragging it to the container of your choice. For each ball you place in the **blue** container, you will receive **2 cents**, and for each ball you place in the yellow container, you will receive **4 cents**.

The rule is to place the balls in the blue container.

Once the task begins, you have a maximum of 10 minutes to put the 50 balls into the containers.

The sum of your payouts from **blue** and yellow container will be added to your payout at the end of the experiment

If you have any questions, please raise your hand. One of the experimenters will then come to your place. When you are ready to start the task, please press "Next".

D.2 Task Page

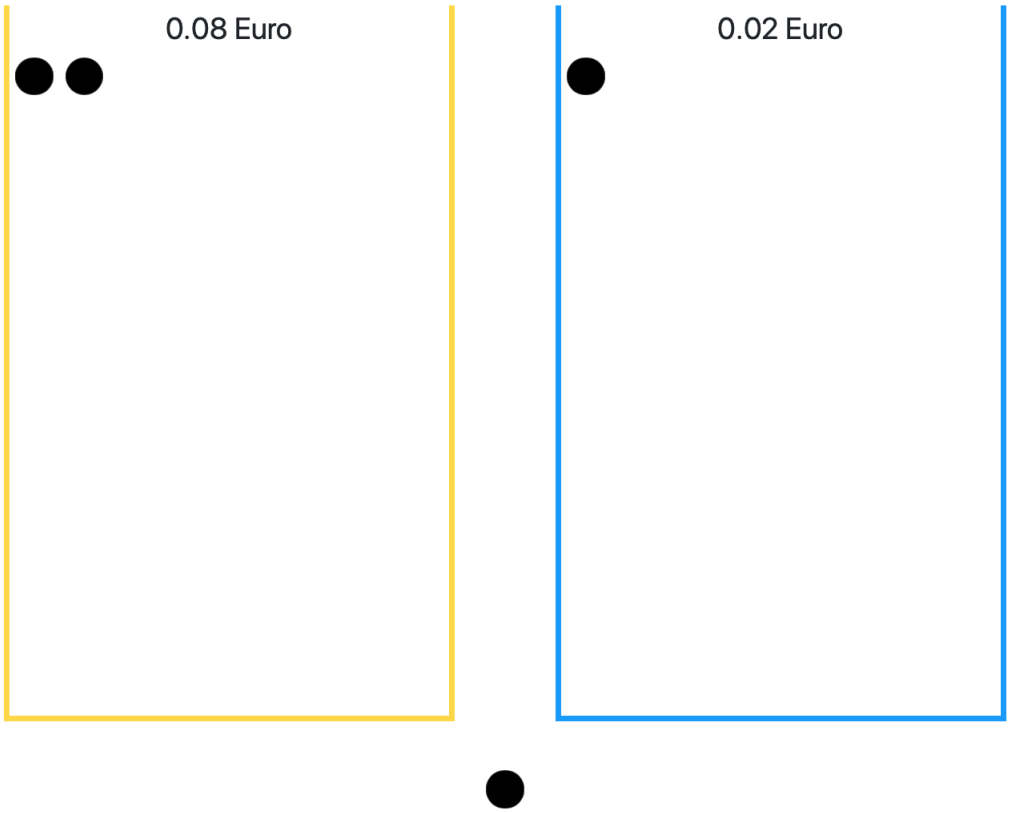

Figure A5: Norm Following Decision Screen