# Social Preferences under the Shadow of the Future*

Felix Kölle† *University of Cologne*

Simone Quercia‡ *University of Verona*

Egon Tripodi§ *Hertie School*

June 2023

**Abstract**

Social interactions predominantly take place under the shadow of the future. Previous literature explains cooperation in indefinitely repeated prisoner's dilemma as predominantly driven by self-interested strategic considerations. This paper provides a causal test of the importance of social preferences for cooperation, varying the composition of interactions to be either homogeneous or heterogeneous in terms of these preferences. Through a series of pre-registered experiments ($N$ = 1,074), we show that groups of prosocial individuals achieve substantially higher levels of cooperation. The cooperation gap between prosocial and selfish groups persists even when the shadow of the future is increased to make cooperation attractive for the selfish and when common knowledge about group composition is removed.

**Keywords**: cooperation, indefinitely repeated games, 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 experiments (#28887, #100729) were pre-registered on the AsPredicted platform. Research funding from the Reinhardt Selten Institute and the Center for Social and Economic Behavior (C-SEB) is gratefully acknowledged. We further gratefully acknowledge the financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 390838866. Support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged. We are especially grateful to Maria Bigoni, Pedro Dal Bo, Guillaume Fréchette, David K. Levine, Daniele Nosenzo, and Andis Sofianos for very constructive feedback. We thank Clara Barrocu, Aenne Läufer, Carina Lenze, Margarita Radkova, Vincent Selz, Valerie Stottuth, and Rosa Wolf 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

§Hertie School, Friedrichstraße 180, 10117 Berlin, Email: tripodi@hertie-school.org.

# 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 raises continued interest across the social sciences. The individual interest that can determine cooperation is the threat of future punishment; the Folk theorem identifies sharp conditions on how large the shadow of the future needs to loom for cooperation to be possible (Fudenberg and Maskin, 1986). Despite being broadly in line with this theory, the empirical evidence is more nuanced and substantial variation in cooperation remains unexplained (see Dal Bó and Fréchette, 2018, for a recent overview).

One natural candidate to explain this variation is the heterogeneity in people's social motivations, which is pervasive across different demographics, cultures, and societies (Henrich et al., 2001; Croson and Gneezy, 2009; Fisman et al., 2015; Falk et al., 2018). Social preferences have further been shown to explain behavioral patterns across a large variety of contexts, both in the lab (Cooper and Kagel, 2016) and in the field (Fehr and Charness, 2023). Yet, despite their widespread influence on behavior, social preferences are surprisingly perceived as unimportant for cooperation in indefinitely repeated games. In a comprehensive review of the literature, Dal Bó and Fréchette (2018, p. 88) write: "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." This is especially noteworthy as studies on one-shot and finitely repeated cooperation games show that many empirical findings can be explained by social preferences (Gintis, 2000; Fischbacher and Gächter, 2010; Fehr and Schurtenberger, 2018).

In this paper we provide a novel test of social preferences as a driver of cooperation in indefinitely repeated games that goes beyond correlating measures of social preferences with cooperation outcomes and thereby overcomes a recurring challenge in this literature. While it is empirically reasonable to expect a positive correlation whenever social preferences *are* a driver of cooperation, two key challenges require attention. First, indefinitely repeated games often produce multiplicity of equilibria such that even self-interested agents may be able to sustain cooperation for strategic motives. Second, composition effects matter: for example, they can induce self-interested players to strategically cooperate if aware of the presence of individuals with social preferences, and they can induce prosocial individuals

who are conditional cooperators to not cooperate if they believe others do not cooperate. To overcome these challenges, our theory-guided experiment is designed to manipulate both the equilibrium outcomes that are attainable depending on the preference composition of participants and to exogenously alter the composition of groups in terms of the prevalence of social preferences. This allows us to take into account both heterogeneity at the individual level and heterogeneity in terms of group composition (which may matter on top of individual heterogeneity as recently shown by Proto et al. (2019, 2022) in the context of intelligence and personality traits).

Our empirical strategy is based on a series of pre-registered laboratory experiments ($N = 1{,}074$) in which we first elicit players' revealed preferences for cooperation using a sequential prisoner's dilemma game. Based on their decisions, we classify individuals as either *selfish* or *prosocial*. Using these elicited preferences, in the second part of our experiment, we sort players into different groups and let them play an indefinitely repeated version of the game. 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 prosocial, and (iii) *mixed groups* that consist of a combination of both prosocial and selfish types.

Our findings reveal that pairing prosocial types with like-minded individuals has a strong and persistent effect on cooperation compared to groups of selfish individuals. We find this result across different experimental conditions. In our main experiment, we consider a situation where (i) participants are informed about the group composition and (ii) the continuation probability, $\delta$, is sufficiently low such that under standard assumptions cooperative equilibria do not exist. These two conditions ensure that, from a theoretical point of view, full cooperation (and many other equilibrium outcomes) can be achieved in groups of prosocial players, while full defection is the only equilibrium outcome among groups of purely self-interested players. In line with this prediction, we find stark differences in cooperation rates across *prosocial groups* (72%) and *selfish groups* (18%).

In two additional experiments, we demonstrate the robustness of this effect when relaxing either condition (i) or (ii). In particular, when relaxing (i) by not announcing the composition of groups, we find similarly strong differences as in our main experiment with *prosocial groups* achieving a cooperation rate of 68% compared to 22% in *selfish groups*. Likewise, when relaxing condition (ii), by considering a situation in which $\delta$ is increased up to a level at which mutual cooperation can be sustained as an equilibrium outcome even among self-interested players, we still find pronounced differences in cooperation across groups. While

*selfish groups* now achieve considerably higher cooperation rates of 40%, they only cooperate half as often as *prosocial groups* (85%).

To test the validity of our measure of social preferences, as elicited in the first part of our experiment, we link it to a series of individual characteristics that previous literature has associated with (preferences for) cooperation. Our results reveal a strong association between our type classification 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; Kölle and Quercia, 2021). In line with previous evidence, we further 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 find no differences across types with regard to risk attitudes or intelligence, ruling out alternative 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; Ghidoni and Suetens, 2022). Most related to our paper are two studies by Dreber et al. (2014) and Davis et al. (2016) who investigate the role of social preferences in indefinitely repeated games by correlating cooperation behavior with donation decisions elicited ex-post. The evidence provided in these papers is mixed, leading Dreber et al. (2014) to conclude that altruism does not play a major role in explaining heterogeneity of play in repeated games. Further related is a study by Reuben and Suetens (2012) who use a sequential version of the repeated prisoner's dilemma in which both players can condition their choice on whether the current round of the interaction is the last one or not. They find strong *end-game* effects that point to cooperation being strategically motivated. Our paper is also related to a recent paper by Kartal and Müller (2022) who propose a model with heterogeneous preferences for cooperation. They show that even when preferences are private information, there exist Bayesian Nash equilibria in which people with strong social preferences play cooperative strategies while other players defect. In such situation, the level of cooperation in equilibrium depends on the composition of interactions, which is in line with our findings. An additional indication that group composition could matter for cooperation comes from existing evidence that being able to choose a partner who cooperates in a repeated gift-exchange game improves cooperation (Bernard et al., 2018). Relative to the existing work, a key contribution of our paper is to *causally* identify the

role of social preferences by exogenously manipulating the composition of groups. As our results demonstrate, group composition effects are crucial for social preferences to matter in repeated contexts, which can explain why previous studies have found no conclusive evidence on the importance of social preferences. Our findings have implications for the formation of social groups and the design of institutions to foster efficiency (see Section 6 for a discussion).

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. Sections 4 and 5 report our experimental results. Section 6 concludes.

# 2 Theoretical Considerations

Consider the monetary payoff matrix represented in Panel (a) of Table 1 for a Prisoner's Dilemma game with $T > R > P > S$. Assuming pure 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) Monetary 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, continue to cooperate until the other player defects, and then defect forever, choosing grim trigger yields a higher monetary 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 monetary 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 showing that many people are not solely motivated by their own monetary payoffs, but that they also care about the well-being of others (see Fehr and Fischbacher, 2003; Sobel, 2005; Cooper and Kagel, 2016, for overviews of the literature). For example, previous literature has shown that many people are willing to share money with strangers (Engel, 2011), help by donating blood or money to charities (Andreoni and Payne, 2013; Slonim et al., 2014), and cooperate even in anonymous one-shot games (Mengel, 2017). Several models of other-regarding 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., $U(\textit{Defect}, \textit{Cooperate}) > U(\textit{Cooperate}, \textit{Cooperate})$ and $U(\textit{Defect}, \textit{Defect}) > U(\textit{Cooperate}, \textit{Defect})$, agents with sufficiently strong social preferences may prefer $U(\textit{Cooperate}, \textit{Cooperate})$ over $U(\textit{Defect}, \textit{Cooperate})$, 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 (see also Brunner et al., 2021). Thus, under the assumption of common knowledge of (sufficiently strong) social preferences, mutual cooperation is (on top of mutual defection) a possible equilibrium outcome of the stage game. As a consequence, mutual cooperation is also an equilibrium outcome in *finitely* as well as in *infinitely repeated* games.[1]

[1] 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 only very few people are willing to cooperate unconditionally. Instead, the large majority of people are only willing to cooperate conditionally on others cooperating as well (Fischbacher et al., 2001; Chaudhuri, 2011; Gächter et al., 2017), a finding we replicate in our experiment (see Section 4).

To test for the importance of social preferences in infinitely repeated games, we design a controlled laboratory experiment (see below). In our empirical strategy, we will remain agnostic about the exact social preference motive at play — which likely is different across individuals — but instead elicit a revealed preference measure for cooperation. The advantage of this approach is that we don't have to assume any specific utility function. Therefore, we can simply represent these preferences with a general utility function $u_i(R, T, P, S, \beta_i)$, where $\beta_i$ is the parameter governing social preferences. Our design will reveal which participants display a $\beta_i$ high enough such that they prefer to cooperate in case the other player cooperates, and which participants have a sufficiently low $\beta_i$ such that defection remains the dominant strategy. For the former participants, the game in utilities becomes a coordination game while for the latter it remains a Prisoner's Dilemma.

# 3 The Experiment

Our experiment consists of three parts. In the first two parts, participants 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 a sequential one-shot version of the game. In part 2, participants play an indefinitely repeated version of the 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.

Table 2: Monetary payoffs in the Prisoner's Dilemma game

| | | **Player 2** | |
|---|---|---|---|
| | | *Cooperate* | *Defect* |
| **Player 1** | *Cooperate* | 15€, 15€ | 0€, 25€ |
| | *Defect* | 25€, 0€ | 10€, 10€ |

## 3.1 Experimental Design

**Part 1: Eliciting preferences for cooperation.** In part 1, participants play a sequential one-shot version of the Prisoner's Dilemma game as shown in Table 2. 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, participants are asked to make one *unconditional* and two *conditional* decisions. In the unconditional (first-mover) decision, participants are simply asked to choose one of the two options, cooperate or defect. In the conditional (second-mover) decisions, participants 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, the unconditional decision is implemented, while for the second-mover the corresponding conditional decision (depending on the first-mover's unconditional decision) is implemented.

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 participants'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.

An important feature of our design is that we elicit participants' other-regarding concerns in Part 1 in a task that is closely connected to the strategic decision situation that participants face in the repeated game. The main reason for this approach is to obtain a revealed preference measure of the willingness to cooperate given others' cooperation (see Section 2), which allows us to separate the participants for whom the game in utilities is a coordination game (*prosocial*) from those for whom the game is a Prisoner's Dilemma (*selfish*). This parsimonious measure can be directly used to predict participants' behavior in the repeated game. Participants classified as cooperators in Part 1 should cooperate in a simultaneous decision in the repeated game if they believe that their counterpart would cooperate. Participants classified as selfish should instead defect regardless of their belief

about other's cooperation. This tight prediction would not be possible if we had chosen a different measure of social preferences in Part 1, or at least would have required structural assumptions and measurements to capture different prosocial motives such as e.g. altruism, inequity aversion or reciprocity.[2]

**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 $\delta$, 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, 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. Participants play a total of twenty supergames. Participants remain matched with the same counterpart for all rounds within a supergame, but are randomly re-matched (in matching groups of ten participants) 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 groups participants are assigned to. We create three different types of groups: 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 participants who in their conditional decision chose to cooperate if the other player cooperates; selfish groups consist only of participants who chose to defect if the other player cooperates; mixed groups consist of a combination of both these types.

In the light of our discussion in Section 2 highlighting the importance of common knowledge of social preferences, before the beginning of the repeated game, participants were informed about the type of group they were assigned to. In particular, at the beginning of part 2 we communicated the exact choice that was used to generate the matching. In prosocial groups, we explained that all participants in the matching group chose to cooperate in the conditional decision in case their counterpart chose to cooperate, while in selfish groups participants were told that everyone chose to defect in this case. In 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] Importantly, to

[2]This challenge is also empirically grounded in experiments included in Proto et al. (2019), where the authors sort players by correlates of cooperative attitudes such as agreeableness and conscientiousness, as they show that these group composition effects on cooperation are small and transitory.

[3]The exact wording for the typical session with 30 participants was as follows: "In each sequence you are paired with a participant randomly drawn from a group of 9 people. The group of 10 people (including

avoid that participants strategically distort their revealed social preferences in part 1 of the experiment, participants were informed about the details of part 2 and the type of group they were assigned to only before the start of the repeated game.

A critical issue when performing experiments where participants are matched according to their behavior in previous parts of the experiment is the possibility of deceiving participants by withholding potentially payoff-relevant information. To tackle this issue, in the instructions of part 1 we informed participants that we would use their decisions for part 2. In particular, we told participants 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 participants were always either placed in a segregated or 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 would have a strong effect on participants' responses in part 1, e.g., if many selfish individuals 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.

**Part 3: Eliciting individual characteristics.** Given recent evidence highlighting the importance of personal characteristics for outcomes in repeated interactions (Proto et al., 2019, 2022), and to provide insights into the "behavioral validity" of our social preference measure elicited in part 1, 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 implement a big-five personality inventory (Schupp and Gerlitz, 2008) and elicit a measure of IQ using a ten-item Raven's progressive matrices test (Raven, 2000). We further elicit participants' general risk attitudes (Dohmen et al., 2011), gender, and their propensity to follow norms. Eliciting norm-following behavior is interesting, because (conditional) cooperative behavior has long been argued to be associated with (the willingness to follow) social norms (see, e.g., Camerer and Fehr, 2004; Fehr and Fischbacher, 2004; Fehr and Schurtenberger, 2018; Kölle and Quer-

yourself) has been determined according to one of the conditional decisions of Part 1." For prosocial [selfish] groups the instructions continued as "In particular, all participants in your group (including you) chose to play A [B] in case the other player chooses A.". For mixed groups the instructions continued as "In particular, some participants in your group chose A in case the other player choose A and some chose B in case the other player chooses A." (see Appendix D for the full instructions). For sessions with fewer than 30 participants, we kept the group size constant at one third of the session size and adjusted the wording of the instructions accordingly.

cia, 2021). To 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 E for the instructions and further details). 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 in selfish groups. Recall that in these groups, we match together 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 ($\beta_i$). As discussed in Section 2, when $\delta < \hat{\delta}^{SPE}$, the unique equilibrium outcome among such self-interested individuals is full defection. As a result, we expect very low levels of cooperation in these groups, especially as participants 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) participants 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. Hence, we expect that some groups of cooperators manage to achieve these cooperative equilibria. This leads us to our second hypothesis:

**Hypothesis 2.** *Prosocial groups 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 (2022), 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 exist 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. In contrast

with prosocial groups, the set of equilibria is less cooperative as mutual cooperation is not feasible. Additional insights regarding our mixed groups can be derived from previous studies on indefinitely repeated games, which have paired participants 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 Additional Experiments

Further to our main experiment, we provide complementary tests of our hypotheses in four additional experiments.

*Robustness experiment: Random matching.* This experiment closely resembles our main experiment except that after the elicitation of preferences (in part 1), participants are matched at random in three equally sized matching groups. This experiment serves two main purposes. First, it allows us to assess whether behavior in groups in which players are told that the composition of types is mixed is comparable to behavior in groups in which players are simply matched at random without further announcements, as done in most previous studies. Second, it allows us to study how naturally occurring variation in the share of prosocial types in groups affects cooperation.

*Robustness experiment: Long-run.* This experiment also resembles our main experiment except that in part 2, participants interact for 40 rather than 20 supergames. This experiment gives even greater scope for the behavior of participants to converge to their equilibrium strategies and allows us to study the persistence of group composition effects.

*Additional experiment: Unannounced matching.* In our main experiment participants are (i) matched into segregated groups and (ii) made aware of the group composition. This way we can fix beliefs about the distribution of types and obtain sharp theoretical predictions on the set of achievable equilibria (see Section 2). However, announcing group composition may also come with some disadvantages. Specifically, announcing the group composition may affect behavior via channels different from (social) preferences. For instance, the announcement may shift beliefs about others' actions which could help prosocial groups to coordinate on high levels of cooperation.[4] Moreover, perfect knowledge of group composition may have

[4]Alternatively, the announcement may create a positive group identity that can arise when being told

little relevance empirically, as in real world settings this information is often unavailable (although people may form some impressions of others based on (imperfect) signals). To control for these issues, we ran an additional experiment in which we diverge from our main experiment only in that we make no mention of how groups are formed (see Sabater-Grande and Georgantzis, 2002; Proto et al., 2019, 2022, for a similar approach). By comparing the results of this experiment to those from our main experiment, we can disentangle the effect of group composition from potential confounding effects of the group composition announcement, and thus also contribute to a recent literature that examines the effects of revealing information about personality data on prosocial behavior (Drouvelis and Georgantzis, 2019; Lambrecht et al., 2022).

*Additional experiment: High delta.* Our main experiment was designed to test stark predictions on cooperation differences between *prosocial* and *selfish* groups. To this end, we relied on a situation in which the game does not have cooperative equilibria for self-interested agents. While such a design helps with identification, at the same time, it studies the importance of social preferences only under restrictive conditions. Therefore, to test for the role of social preferences more broadly, in an additional experiment we study a situation in which mutual cooperation is a subgame-perfect Nash equilibrium of the game even for self-interested agents. Studying such a situation is interesting not only because it is the one most often studied in previous literature (Dal Bó and Fréchette, 2018, p. 74), but also because social preferences are no longer necessary to sustain cooperation. Evidence that they do matter even in such contexts would support the view that social preferences matter for cooperation more generally. The design of this additional experiment is identical to our main experiment, except that in part 2 we increase the continuation probability to $\delta = 0.8 > \hat{\delta}^{SPE} = 0.67$.[5]

## 3.4 Procedures

The sessions of our experiments were conducted at the Cologne Laboratory for Economic Research (CLER) at the University of Cologne and the Decision Lab at the Max Planck

to be "playing with the good guys", or affect cooperation through social norms whereby selfish types don't feel bad by defecting because it is announced that the predominant social norm in their group is to defect.

[5]At such $\delta$, both mutual cooperation and mutual defection are possible equilibria for selfish, mixed, and prosocial groups. 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 them to also want to cooperate.

Institute for Research on Collective Goods in Bonn. In total, we ran thirty-eight sessions with $N = 1,074$ participants (see Table A1 in Appendix A for an overview). Participants were students from various disciplines recruited from the subject pools of the universities in Bonn and Cologne. Our main experiment and our additional experiments were pre-registered on the AsPredicted platform (#18979, #100729 and #28887).

At the beginning of each session, participants 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 D for an English copy of the instructions). After that, participants read the instructions for part 1 of the experiment, followed by control questions designed to ensure participants' understanding of the game. Only after each participant answered all the questions correctly, part 1 started. Upon completion of part 1, participants received instructions for part 2. The instructions were again followed by a set of control question, testing participants' understanding. As before, the experiment continued only after all the questions were answered correctly by each participant. In our main experiment as well as the long-run and the high delta experiments, before the beginning of part 2 participants were informed about the type of group they were assigned to. In the random matching and the unannounced matching experiments, in contrast, participants were simply told that they interact in fixed matching groups but nothing was said about the composition of groups. 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. For $\delta = 0.6$, the length of the supergames ranged between 1 and 13 rounds with an average of 2.39. For $\delta = 0.8$, the length of the supergames ranged between 1 and 24 rounds, with an average of 4.93.

After finishing part 2, but before learning about their earnings from the experiment, participants 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 participants' earnings. If part 1 was selected, participants were paid either according to their unconditional or their conditional decision as described above. If part 2 was selected, the computer randomly selected one supergame and paid the last round of that supergame. As shown by Sherstyuk et al. (2013), paying only for the last instead of all rounds has the advantage that it is theoretically robust to different risk attitudes (see also Dal Bó and Fréchette, 2018, for a further discussion of different payment methods). We further 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 participants to distort their choices in part 1. In addition, participants received their earnings from the norm-following task, ranging between 1 and 2 Euro. On average, participants earned 17.27 Euros for sessions that lasted between one and one and a half hours.

# 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. After that, we describe how grouping different types of participants into mixed and segregated groups affects cooperation in the indefinitely repeated game in part 2.

## 4.1 Cooperation types

In the unconditional decision of part 1, 37% of the participants chose to cooperate. In the conditional decisions, cooperation rates amount to 7% when choosing conditional on the partner's choice to defect and to 44% when choosing conditional on the partner's choice to cooperate. As explained in Section 3.1, we use the latter choice to distinguish between participants with high and low degrees of social preferences. Based on this, we classify 44% of participants as prosocial and 56% selfish. Note that given that almost all our participants (93%) chose to defect conditional on other's defection, we find that almost all of the participants that 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 participants that we classify as selfish are free-riders who choose to defect irrespective of the other player's choice (see Table A2 in Appendix A for a full breakdown of choices).[6] Using this classification, we then formed the different groups as described above.[7]

[6]When we use both conditional choices to classify people into types as in previous studies, we find 4% unconditional cooperators, 40% conditional cooperators, 53% unconditional defectors (free-riders), and 3% mis-matchers. The distribution of types is thereby very similar to the ones reported by Nosenzo and Tufano (2017) and Miettinen et al. (2020), for slightly different parameters of the stage game.

[7]Note that because types were determined endogenously within each session, it was not always possible to form all three types of groups. In particular, in total we had three sessions without a prosocial group and four sessions without a mixed group, leading to a total of 32 prosocial groups, 31 mixed groups, and 42 selfish groups. In addition, we had nine groups with complete random matching. Further note that due to some no-shows, some sessions were run with fewer than thirty participants.

## 4.2 The effects of social preferences on cooperation when group composition is known

Using the data from our main experiment, we provide a first test our three hypotheses from the previous section. As pre-registered, in our analyses we mainly focus on first round cooperation because supergames may have different lengths and cooperation rates may depend on histories within these supergames. We show that all our findings are robust to using data from further rounds.

Figure 1 gives a complete overview of the main findings. The left panel shows, separately for each group type, the evolution of first round cooperation rates across supergames while the right panel shows total averages over all supergames (see Figure A1 in Appendix A for a breakdown of cooperation rates by matching group). We observe pronounced differences in cooperation rates across the three types of groups. 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 cooperation 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.[8] 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 Table C1 in Appendix C). This indicates that the differences in cooperation we observe do not only arise at the beginning of a supergame, but are maintained through repeated interactions.

Mixed groups, which on average are composed by 50% prosocial types and 50% selfish types, start off with intermediate levels of cooperation that are higher than in selfish groups

[8]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. We tackle this question in two ways. First, we follow the approach by Kartal and Müller (2022) who, based on a technique proposed by Noussair et al. (1995) and Barut et al. (2002), estimate asymptotes of cooperation rates. The results from 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). Second, we can rely on the data from our long-run experiment (see Section 3.3) in which participants interacted for forty instead of twenty supergames. The results, shown in Figure A2 and Table A6 in Appendix A, reveal that cooperation rates in prosocial groups remain significantly higher than the ones observed in selfish groups even after having interacted for twenty supergames. Average first round cooperation rates in supergames 21-40 amount to 0.61 in prosocial groups and 0.22 in selfish groups, a difference that is highly significant ($p = 0.006$). Similar results hold when looking at data from all rounds (0.49 vs. 0.16, $p = 0.015$).

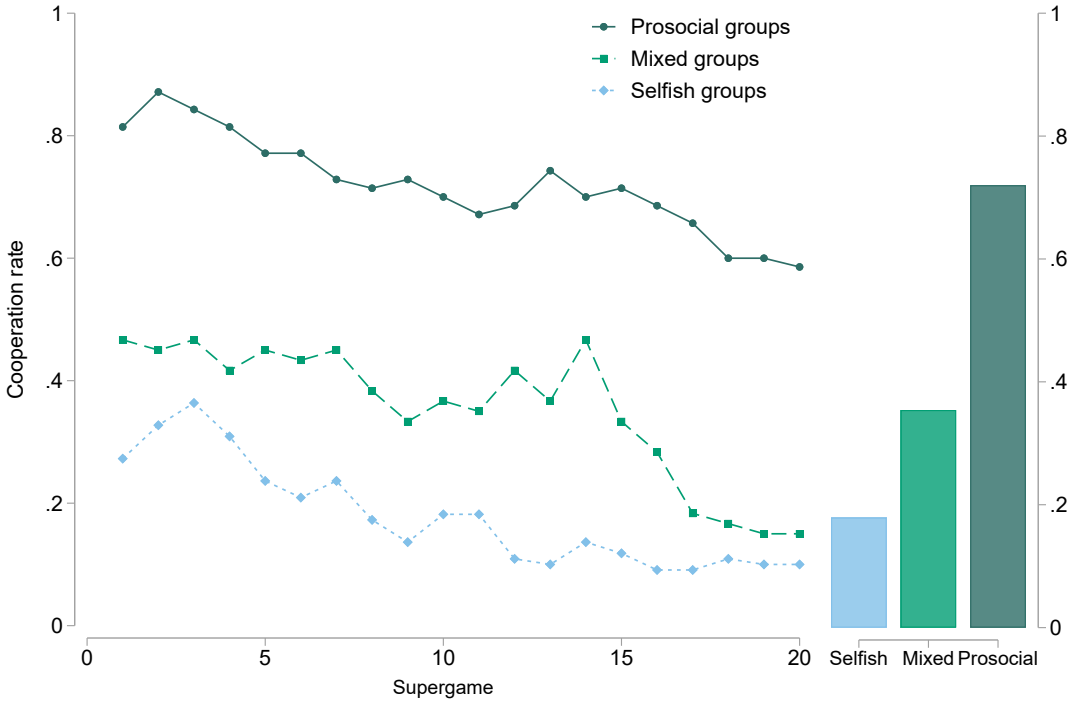


Figure 1: First round cooperation by group type

but lower than in prosocial groups. As the game proceeds, however, cooperation rates decrease to similar levels as in selfish groups. Averaged over all supergames first round cooperation rates amount to 35%, which is only slightly higher than the levels observed in selfish groups, but much lower than in prosocial groups. This is in line with 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 ten, and all supergames combined, along with the significance levels from each pairwise comparison. Columns 2-4 show cooperation levels using first round data, while Columns 5-7 report the results using data from all rounds.

The results 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. Furthermore, the differences remain statistically significant and similar in size when considering only the last ten supergames or when looking at cooperation rates from all rounds. This is reassuring

Table 3: Cooperation rates across supergames and group type

| | First round | | | All rounds | | |
|---|---|---|---|---|---|---|
| | 1 | 11-20 | All | 1 | 11-20 | All |
| Prosocial groups | 0.81 | 0.66 | 0.72 | 0.73 | 0.48 | 0.58 |
| Selfish groups | 0.27 | 0.11 | 0.18 | 0.21 | 0.09 | 0.14 |
| Mixed groups | 0.47 | 0.29 | 0.35 | 0.41 | 0.20 | 0.29 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p = 0.016$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p < 0.001$ | $p = 0.017$ | $p = 0.010$ | $p = 0.006$ | $p = 0.033$ | $p = 0.042$ |
| $H_0$: Mixed = Selfish | $p = 0.007$ | $p = 0.185$ | $p = 0.144$ | $p < 0.001$ | $p = 0.301$ | $p = 0.208$ |

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

as it indicates that experience or learning effects do not diminish the group composition effects that we observe. 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 ten supergame or all supergames combined. We summarize these findings in our first result:

> **Result 1:** *Grouping prosocial types together in segregated groups has a strong positive effect on cooperation rates.*

To shed some further light on 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, we estimate repeated-game strategies using the Strategy Frequency Estimation Method (SFEM) as proposed by Dal Bó and Fréchette (2011). Based on previous literature (Dal Bó and Fréchette, 2019), we use maximum likelihood to estimate the prevalence of the following five strategies: 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.[9] Table 4 reveals several interesting patterns. First, in line with previous experiments where $\delta$ is not conducive to cooperation (Dal Bó and Fréchette, 2011), we find that in

[9]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

Table 4: Estimated strategy frequencies

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.264*** | 0.679*** | 0.704*** |
| | (0.093) | (0.141) | (0.075) |
| Always cooperate (AC) | 0.091 | 0.033 | 0.003 |
| | (0.071) | (0.031) | (0.013) |
| Grim trigger (GT) | 0.199*** | 0.095 | 0.061 |
| | (0.079) | (0.075) | (0.044) |
| Tit-for-tat (TFT) | 0.395*** | 0.166** | 0.056 |
| | (0.107) | (0.078) | (0.038) |
| Suspicious Tit-for-tat (STFT) | 0.051 | 0.027 | 0.177*** |
| | (0.038) | (0.041) | (0.074) |
| Gamma | 0.483*** | 0.493*** | 0.436*** |
| | (0.052) | (0.055) | (0.056) |
| Frequency of cooperative strategies | 0.736 | 0.321 | 0.296 |
| Observations | 70 | 60 | 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. See Tables C4 and C5 in Appendix C for robustness checks.

mixed and selfish groups AD is the most common strategy, amounting to 68% and 70%, respectively. 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. The fact that prosocial types predominantly play conditionally cooperative strategies is consistent with our findings from the first part of the experiment in which the large majority of prosocial types revealed that they are willing to cooperate only conditionally on others doing so, too.

To put our results into perspective, we can compare the achieved levels of cooperation in the different groups to those observed in previous studies. A recent survey by Dal Bó and Fréchette (2018) reports nine studies with $\delta < \delta^{SPE}$ and a minimum of seven supergames and

improves the efficiency of the estimation and reduces the trembling probability *Gamma*, but that this does not qualitatively affect our point estimates (see Table C4 in Appendix C). As an additional robustness check, we re-run our analysis by excluding the data from the first ten supergames, thus focusing only on data from those supergames in which participants already have gained experience in the game. The results, reported in Table C5 in Appendix C, are qualitatively very similar to the ones reported in the main text.

finds that first round cooperation rates in the seventh supergame vary between 2% and 42%. The results from our selfish and mixed groups broadly fall within that range, amounting to 24% and 45%, 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.

In our analyses above, we have argued that behavior in mixed groups can be taken as a proxy for previous findings from indefinitely repeated experiments. The only difference with regard to the matching is that participants in our experiment were explicitly told that they would interact in mixed groups. To ascertain whether such announcement had any behavioral effect per se, we now turn to the data from our random matching experiment. We find that cooperation rates in this experiment are not significantly different from the ones observed in mixed groups in our main experiment, both when looking at first round cooperation rates ($p = 0.261$) and when using data from all rounds ($p = 0.200$). At the same time, in line with our results above (compare Table 3), we find cooperation to be significantly lower than in prosocial groups (both $p < 0.001$) and not significantly different from selfish groups (both $p > 0.555$). Overall, these results indicate that the results from mixed groups in our main experiment are a good proxy for findings from previous literature that have matched participants at random.

## 4.3 The effects of social preferences on cooperation when group composition is unknown

As a next step in our analysis, we move to the results from our first additional experiment in which group composition was not announced. This allows us to disentangle the effect of group composition from potential confounding factors due to its announcement.

Our main findings are shown in Figure 2, displaying average first round cooperation rates both overall and across supergames (see Figure A3 in Appendix A for disaggregated data at the matching group level). Differences in cooperation across the different group types remain stark as in the main experiment. Prosocial groups achieve the highest levels of cooperation (0.68), followed by mixed (0.46) and selfish groups (0.22). As shown in Table A4 in Appendix A, the differences in cooperation rates are all statistically significant, both when considering data from only first or all rounds. Because differences between groups emerge already in the first supergame and remain stable as the game proceeds, we conclude that learning or experience effects do not mitigate our group composition effects.

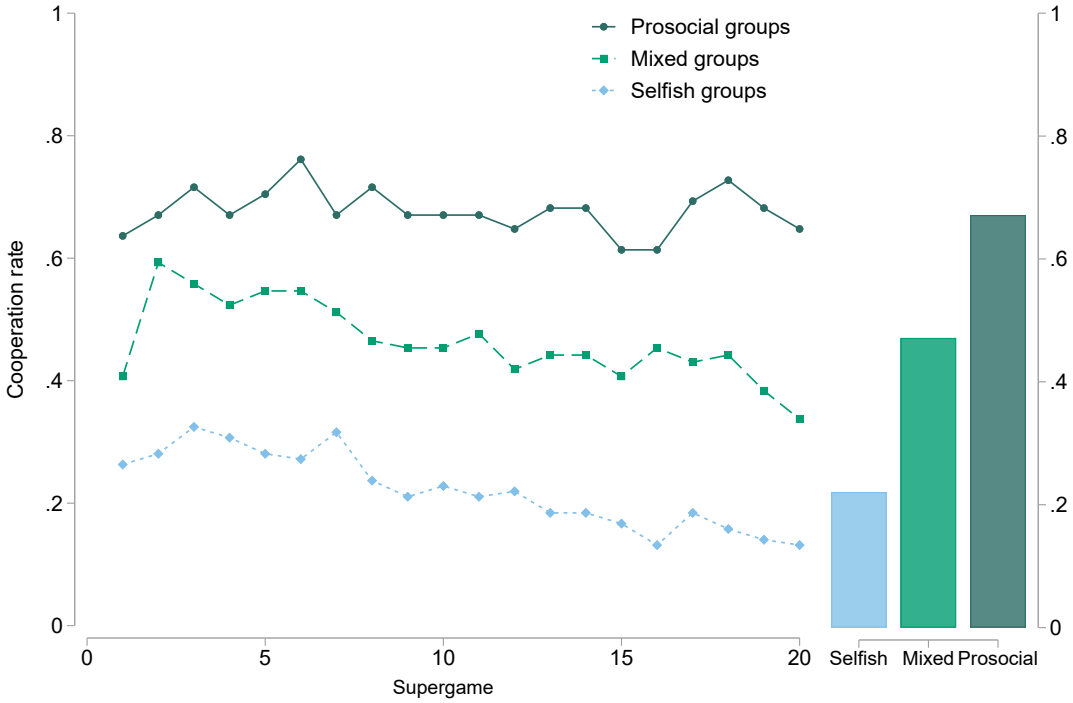


Figure 2: First round cooperation by group type ($\delta = 0.6$, unannounced matching)

In line with the results from our main experiment, estimations on repeated game strategies reveal that while a large majority of decisions in prosocial groups (77%) are best described by cooperative strategies (with TFT being the modal one), the predominant strategy in selfish and mixed groups is always defect (AD) with an estimated frequency of 80% and 42%, respectively (see Table A7 in Appendix A for further details). We summarize these findings in our second result:

> **Result 2:** *Grouping prosocial types together in segregated groups has a strong positive effect on cooperation rates even when the group composition is unknown.*

These results are remarkable because, as laid out in Section 2, segregation of types and common knowledge about it are both necessary theoretical conditions for a cooperative equilibrium to exist in prosocial groups. Here, we show that prosocial groups manage to coordinate on very high levels of cooperation even when the group composition is unknown. Moreover, comparing these results to our main experiment suggests that the announcement per se has a very minor effect on overall levels of cooperation. Specifically, when group composition is known, first round cooperation rates in prosocial groups amount to 0.72, compared

to 0.68 when group composition is not announced ($p = 0.711$). Similar small differences are observed for selfish (0.18 vs. 0.22, $p = 0.516$) and mixed groups (0.35 vs. 0.46, $p = 0.385$). Taken together, these results demonstrate that group composition announcements have negligible effects relative to the effects of group composition itself.

As a final test for the importance of group composition for cooperation, we take a closer look at mixed groups in our random matching and unannounced matching experiments. As in previous experiments in the literature, in both of these settings participants were not informed about how groups were composed. Even though we have relatively few observations, we can use this to test whether the natural variation in the fraction of prosocial individuals in the group had an impact on cooperation. Our data reveals that this is indeed the case. Specifically, we find a strong positive correlation between the fraction of prosocial types and overall cooperation rates (Pearson's correlation, first round: $\rho = 0.41$, $p = 0.088$, all rounds: $\rho = 0.46$, $p = 0.055$, see also Figure A4 in Appendix A). These results are consistent with the model of Kartal and Müller (2022), which predicts that the number of cooperative equilibria increases in the number of cooperators in a group (see, in particular, Proposition 1 (iv)).

## 4.4 The role of social preferences when cooperation can be sustained also among self-interested players

As a final step in our analysis, we move to the results from our second additional experiment in which we set $\delta = 0.8 > \hat{\delta}^{SPE} = 0.67$. Recall from Section 3 that for such high $\delta$, cooperation can be sustained as an equilibrium outcome even among purely self-interested players and that, therefore, we have created a situation in which social preferences are not expected to matter for cooperation outcomes since all three group types share the same set of equilibria.

Aggregate results from this experiment are shown in Figure 3 (see Figure A5 in Appendix A for disaggregated data at the matching group level). Once again, they reveal a strong positive effect of social preferences in promoting cooperation. Averaged over all supergames, cooperation rates amount to 85% in prosocial groups and 40% in selfish groups. As we show in Table A5 in Appendix A, this difference is highly significant, also when considering only the last ten supergames or when using data from all rounds. As before, we find cooperation rates in mixed groups to lie in between these two extremes. Different from the case with $\delta = 0.6$, however, cooperation levels in mixed groups are now closer to the ones observed in prosocial groups, although still significantly lower.

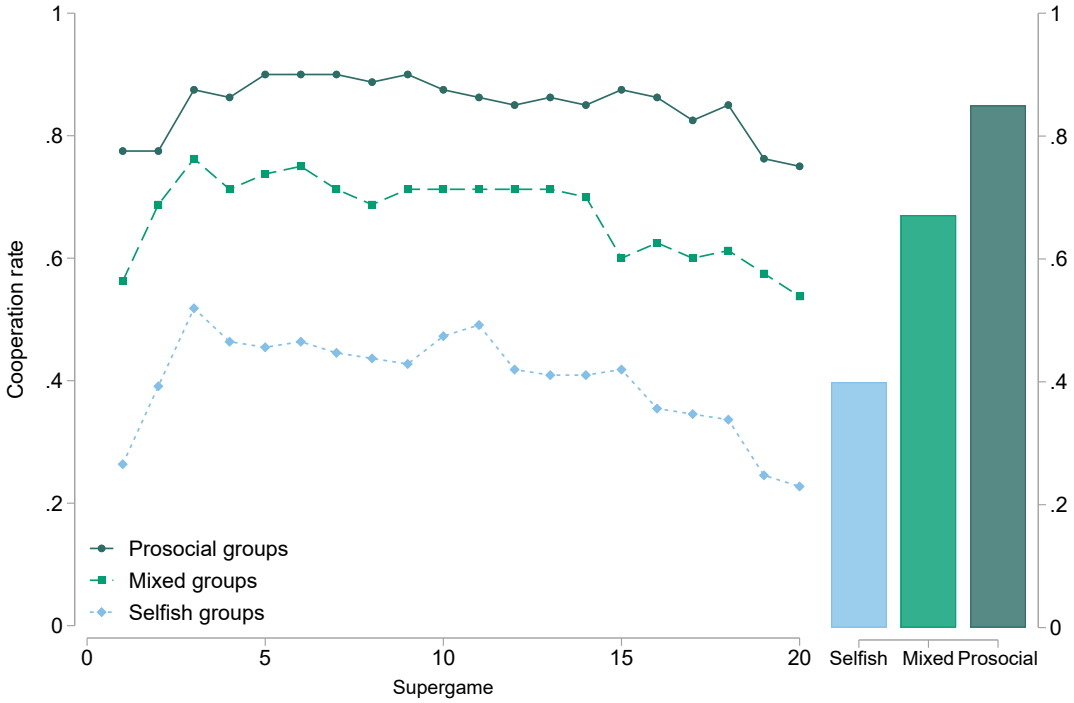


Figure 3: First round cooperation by group type ($\delta = 0.8$)

The results from estimations on repeated game strategies help explain these results. As shown in Table A8 in Appendix A, we find that, despite the fact that cooperation is a possible equilibrium outcomes, in selfish groups cooperative strategies account for only about 51% of chosen strategies. The remaining 49% of choices are best described by AD, which explains why selfish groups largely fail to coordinate on the efficient outcome of mutual cooperation. In mixed and prosocial groups, in contrast, cooperative strategies account for 79% and 90% of choices, respectively (in all groups, the modal cooperative strategy is TFT, followed by GT). This, in turn, explains why these groups are successful in achieving high levels of cooperation.

Comparing these results to the ones from our main experiment indicates that, in line with evidence from previous literature (Dal Bó and Fréchette, 2018), increasing the continuation probability $\delta$ (from 0.6 to 0.8) positively affects cooperation. On average, first round cooperation rates increase from 0.38 to 0.61 ($p = 0.005$). How do the different group types react to the increase in $\delta$? By the Folk Theorem, mutual cooperation now becomes an equilibrium of the game also for mixed and selfish groups. In turn, we should expect an increase in cooperation for these groups. This is not the case for prosocial groups, as for them mutual

cooperation was already achievable for low levels of $\delta$. Our empirical findings align with these predictions, as we observe significant increases in cooperation rates for mixed (from 0.35 to 0.67, $p = 0.028$) and selfish groups (from 0.18 to 0.40, $p = 0.005$), but not for prosocial groups (from 0.72 to 0.85, $p = 0.200$).

Overall, these results reveal that even in situations in which from a theoretical perspective social preferences should not matter for outcomes – in our case selfish, mixed, and prosocial groups exhibit the same set of achievable equilibria – grouping together individuals with prosocial attitudes has a strong positive effect on cooperation. This constitutes our third result:

> **Result 3:** *Even when cooperation can be sustained in equilibrium among self-interested players, a large gap in cooperation remains between prosocial and selfish groups.*

# 5 Validity of our type classification

Our findings above have revealed strong differences in cooperation rates across prosocial and selfish groups, both when $\delta$ is conducive to cooperation and when it is not, and when group composition is known or 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 participants' 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 using data from all $N = 1,074$ participants. We show that our type classification is associated with tendencies to follow norms as well as with personality traits, but does not relate to cognitive ability.

*Norm following.* Kimbrough and Vostroknutov (2016, 2018) have shown that behavior in the norm-following task predicts social behavior in a variety of contexts, including dictator-game giving and second-mover behavior in a trust game. Here, we demonstrate that behavior in this task is strongly correlated to participants' revealed preferences for cooperation as elicited in part 1 of the experiment. This is shown in Figure A6 in Appendix A, depicting the distribution of balls that were put in the prescribed blue urn, separately for participants being classified as prosocial and selfish. The figure reveals pronounced differences 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 23.2 (25) for prosocial and 15.3 (4) for selfish types (t-test, $p < 0.001$). This result demonstrates that cooperation is indeed associated with willingness to follow social norms, as recently argued by Fehr and Schurtenberger (2018) and shown by Kölle and Quercia (2021).

*Personality traits.* We measure personality using the big-five personality inventory (Schupp and Gerlitz, 2008). Previous studies have provided a link between personality traits and cooperativeness (e.g. Becker et al., 2012; Volk et al., 2012; Proto et al., 2019). In line with this previous evidence, we find the following two traits to be most strongly correlated with our type classification: relative to selfish types, prosocial types score higher on agreeableness (t-test, $p = 0.004$) and lower on conscientiousness (t-test, $p = 0.048$).

*Cognitive ability.* Cognitive ability is measured as performance in a 10-item Raven's progressive matrices test (Raven, 2000). Previous literature has highlighted the importance of cognitive skills for a variety of economic primitives 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 reveal no systematic relationship between prosociality and performance in the Raven's test: prosocial types solve on average 4.43 tasks correctly compared to 4.49 for selfish types (t-test, $p = 0.640$; see also Figure A7 in Appendix A). This result indicates that differences in cognitive abilities are unlikely to be the explanation for the different cooperation rates we observe between the different group types.[10]

# 6 Conclusions

Social interactions predominantly take place under the shadow of the future. Previous literature on indefinitely repeated games has emphasized the primary role of self-interested strategic factors in explaining outcomes. In this paper, we present a series of experiments highlight the significance of social preferences in such contexts. Our findings demonstrate

[10]In Table A9 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, while conscientiousness 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 participants' willingness to reciprocate others' cooperation.

that high levels of cooperation can persist despite strategic incentives to defect when prosocial individuals interact in segregated groups. These conclusions hold true across various conditions, including the announcement or non-announcement of group composition, as well as favorable or unfavorable continuation probabilities for cooperation. Our results are important because they provide novel insights into the role of other-regarding motivations in repeated interactions, which can help explain some of the unexplained variation in cooperative behavior observed in previous literature (Dal Bó and Fréchette, 2018).

Our study also offers more general insights into the importance of a common level of prosociality in groups, highlighting its implications for the success of organizations and societies as a whole. One direct implication is that, particularly in situations with a short-term perspective, organizations can benefit from creating homogeneous groups in terms of social preferences. Specifically, based on the data from our main experiment, our results indicate that forming two segregated groups (one comprising prosocial individuals and the other comprising selfish individuals) instead of two mixed groups can enhance group performance in terms of achieved cooperation by 30 percent (from 0.35 to 0.45, compare Table 3). We note, however, that the effectiveness of this approach depends on various factors such as the length of interaction as well as people's information about their peers' types, as shown by our additional experiments. Our findings further emphasize the value of attracting prosocial individuals and fostering a culture of prosociality within organizations. For instance, when recruiting new employees, firms can utilize screening devices, such as socially beneficial commitments, to create incentives for individuals to self-sort themselves into different groups (Brekke et al., 2011; Grimm and Mengel, 2009; Hauge et al., 2019). Additionally, organizations can employ other strategies to encourage prosocial behavior among existing groups. These may include 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).

# 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.

Andreoni, J. and A. A. Payne (2013): "Charitable giving," in *Handbook of public economics*, Elsevier, vol. 5, 1–50.

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.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012): "The relationship between economic preferences and psychological personality measures," *Annu. Rev. Econ.*, 4, 453–478.

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.

Bernard, M., J. Fanning, and S. Yuksel (2018): "Finding cooperators: Sorting through repeated interaction," *Journal of Economic Behavior & Organization*, 147, 76–94.

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.

Brunner, C., T. F. Kauffeldt, and H. Rau (2021): "Does mutual knowledge of preferences lead to more Nash equilibrium play? Experimental evidence," *European Economic Review*, 135, 103735.

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.

Camerer, C. F. and E. Fehr (2004): "Measuring social norms and preferences using experimental games: A guide for social scientists," *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 97, 55–95.

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.

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

Croson, R. and U. Gneezy (2009): "Gender differences in preferences," *Journal of Economic literature*, 47, 448–474.

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.

Drouvelis, M. and N. Georgantzis (2019): "Does revealing personality data affect prosocial behaviour?" *Journal of Economic Behavior & Organization*, 159, 409–420.

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

Engel, C. (2011): "Dictator games: A meta study," *Experimental economics*, 14, 583–610.

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

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018): "Global evidence on economic preferences," *The Quarterly Journal of Economics*, 133, 1645–1692.

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

Fehr, E. and G. Charness (2023): "Social preferences: fundamental characteristics and economic consequences," *CESifo Working Paper No. 10488*.

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

——— (2004): "Social norms and human cooperation," *Trends in Cognitive Sciences*, 8, 185–190.

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

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.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits (2015): "The distributional preferences of an elite," *Science*, 349, aab0096.

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.

Ghidoni, R. and S. Suetens (2022): "The effect of sequentiality on cooperation in repeated games," *American Economic Journal: Microeconomics*, 14, 58–77.

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.

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.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001): "In search of homo economicus: behavioral experiments in 15 small-scale societies," *American Economic Review*, 91, 73–78.

Kartal, M. and W. Müller (2022): "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.

Kölle, F. and S. Quercia (2021): "The influence of empirical and normative expectations on cooperation," *Journal of Economic Behavior & Organization*, 190, 691–703.

Lambrecht, M., E. Proto, A. Rustichini, and A. Sofianos (2022): "Intelligence Disclosure and Cooperation in Repeated Interactions," Tech. rep., IZA Discussion Papers.

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

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, 1351–1390.

——— (2022): "Intelligence, Errors, and Cooperation in Repeated Interactions," *The Review of Economic Studies*, 89, 2723–2767.

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.

Sabater-Grande, G. and N. Georgantzis (2002): "Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test," *Journal of Economic Behavior & Organization*, 48, 37–50.

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.

Sherstyuk, K., N. Tarui, and T. Saijo (2013): "Payment schemes in infinite-horizon experimental games," *Experimental Economics*, 16, 125–153.

Slonim, R., C. Wang, and E. Garbarino (2014): "The market for blood," *Journal of Economic Perspectives*, 28, 177–196.

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

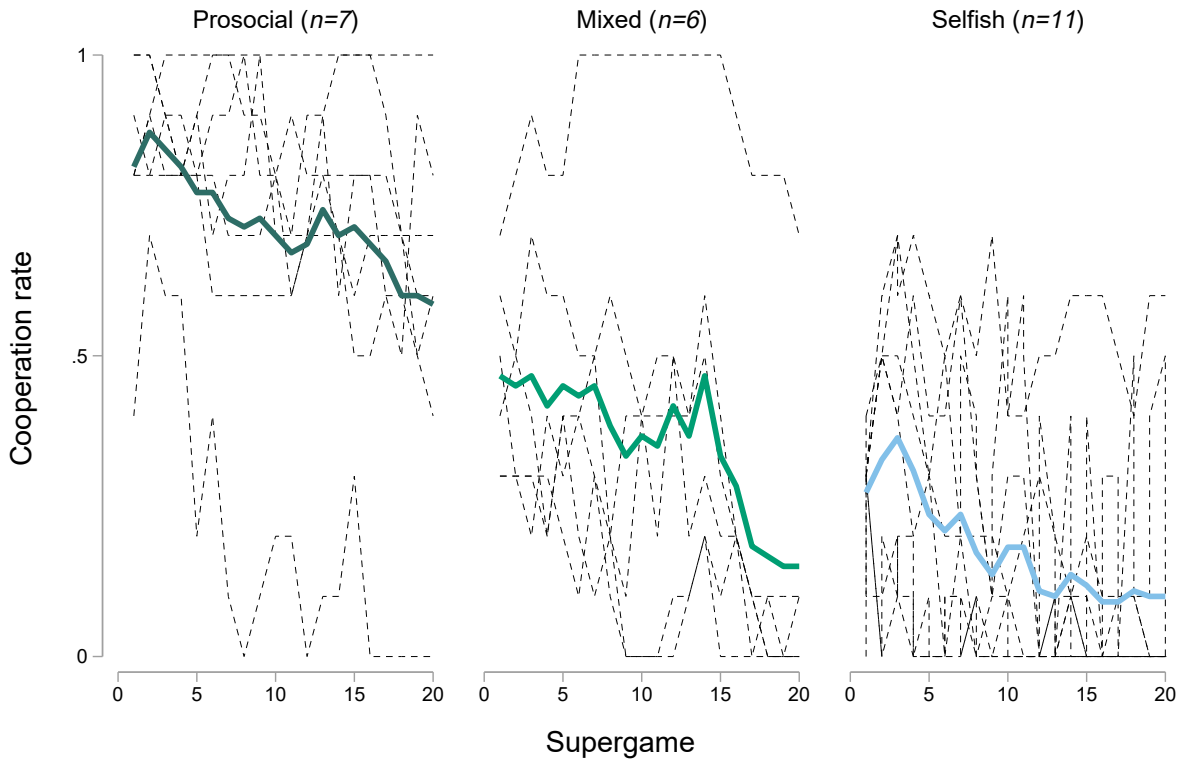


Figure A1: First round cooperation for $\delta = 0.6$ by matching groups when group composition is known. Thick lines display overall averages.

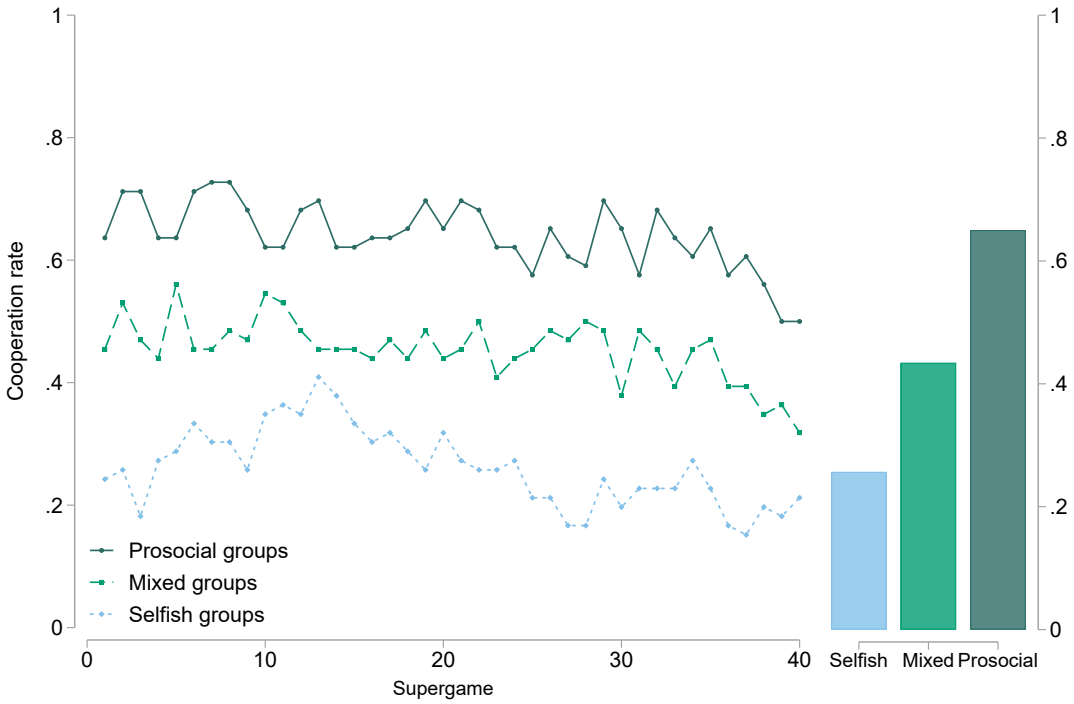


Figure A2: First round long-run cooperation by group type when group composition is known and $\delta = 0.6$

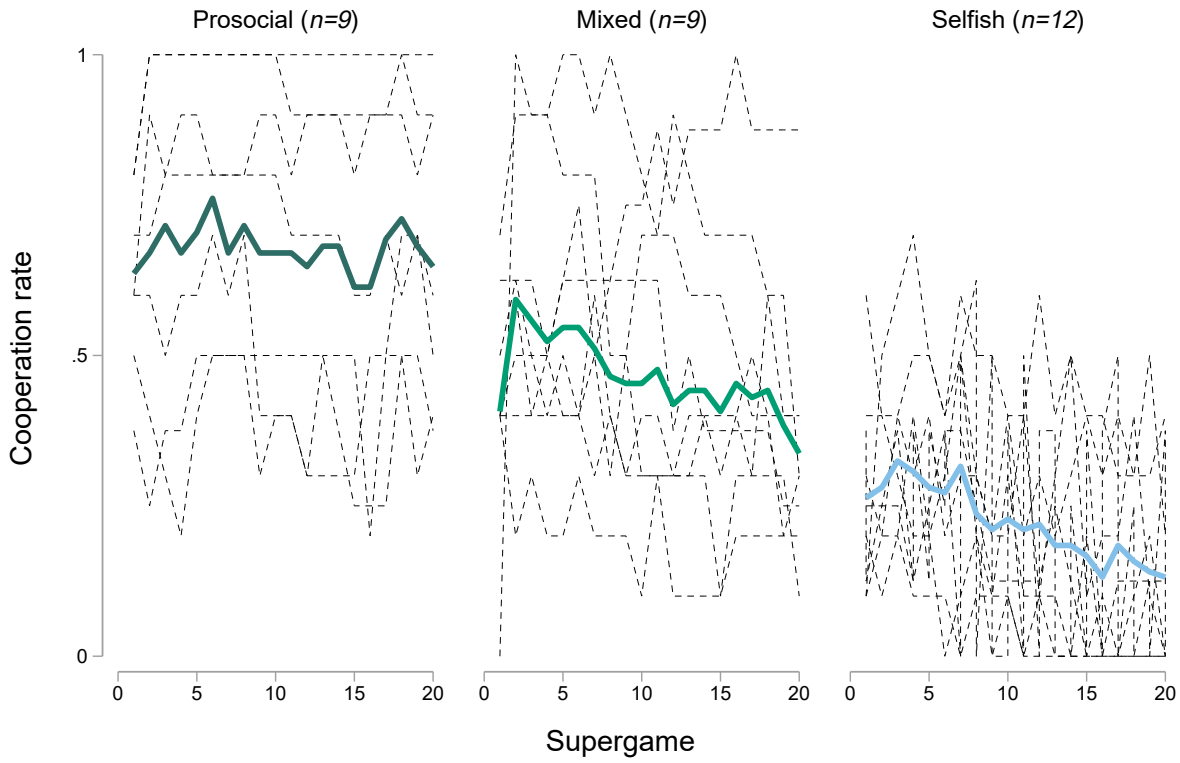


Figure A3: First round cooperation for $\delta = 0.6$ by matching groups when group composition is unknown. Thick lines display overall averages.

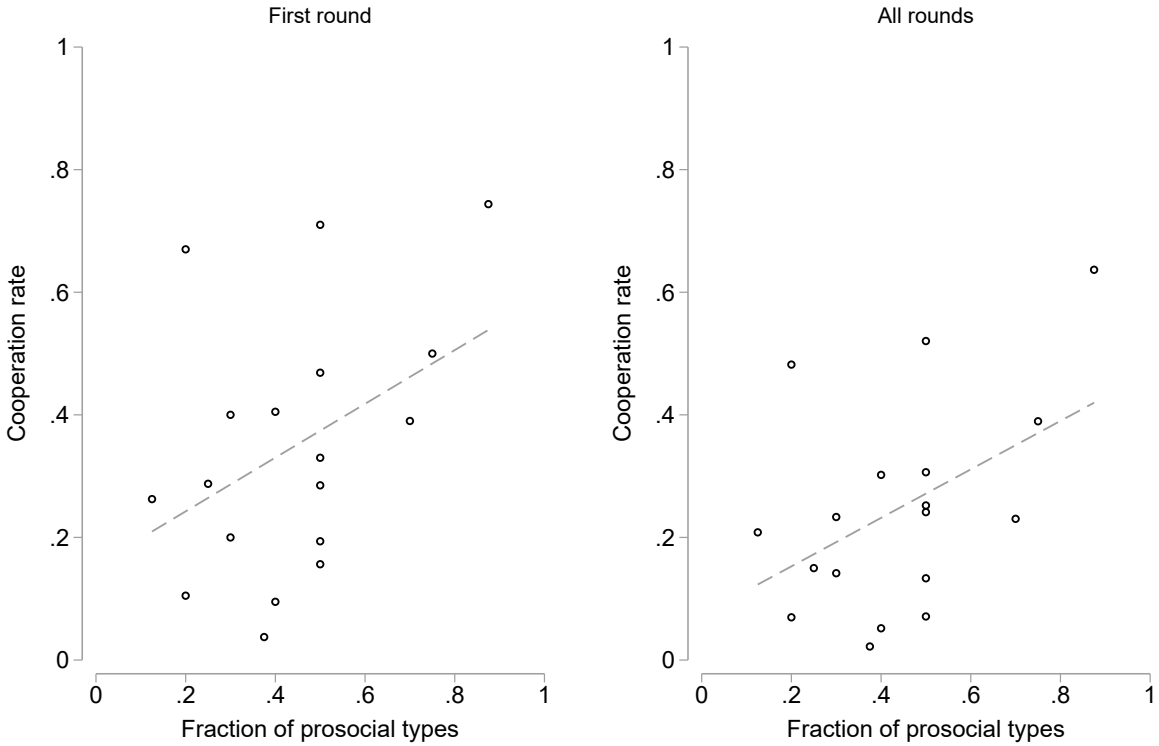


Figure A4: Average cooperation rates in mixed groups by the fraction of prosocial types. Each dot corresponds to one matching group. Left panel: first round cooperation. Right panel: All rounds cooperation.

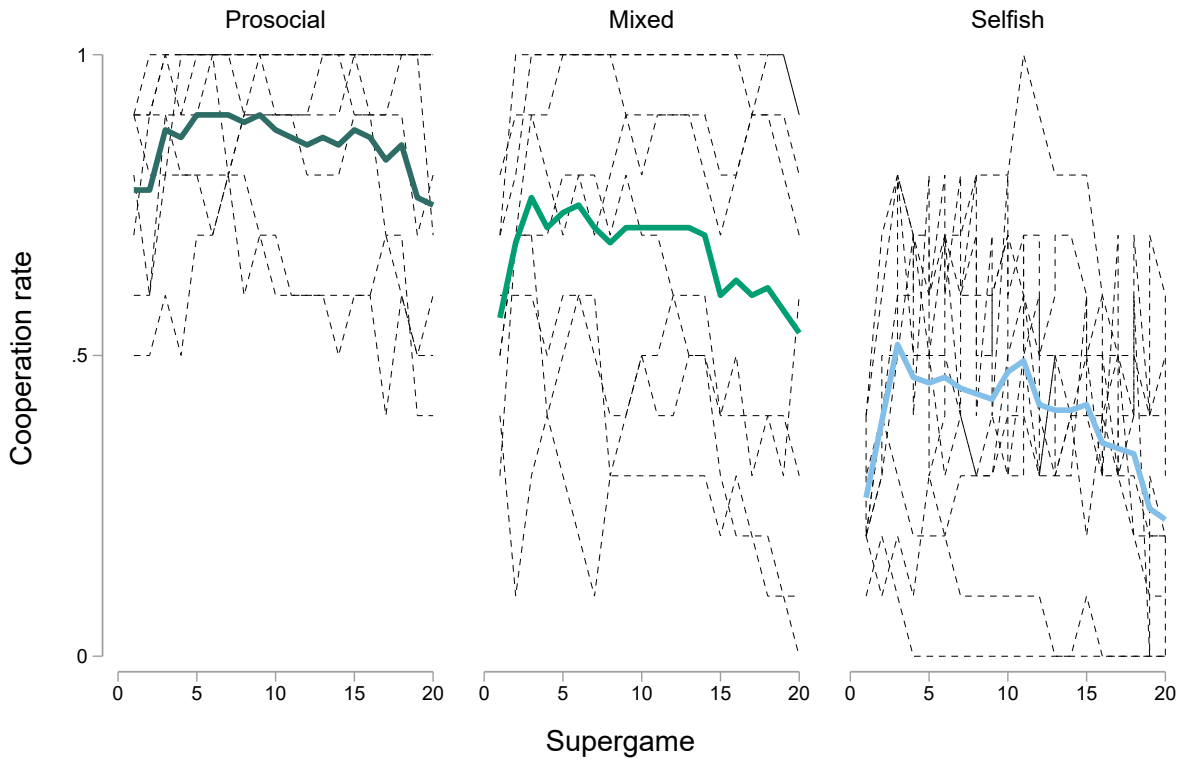


Figure A5: First round cooperation for $\delta = 0.8$ by matching groups when group composition is known. Thick lines display overall averages.

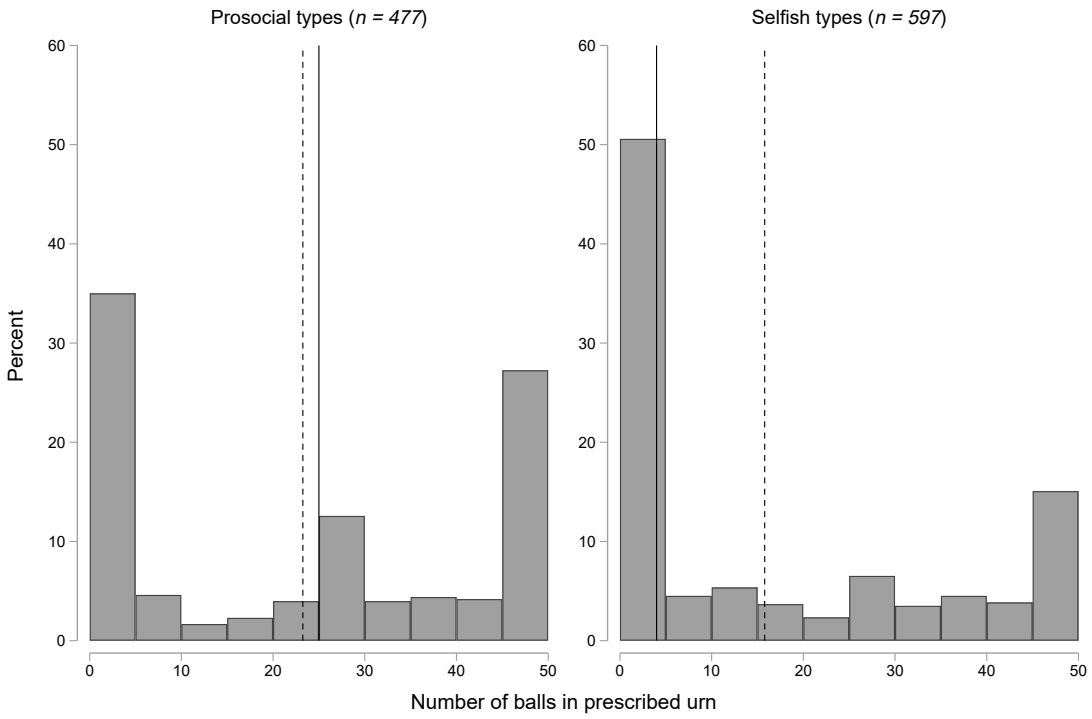


Figure A6: Distribution of the willingness to follow norms. Dashed lines display mean scores, solid lines display the median.

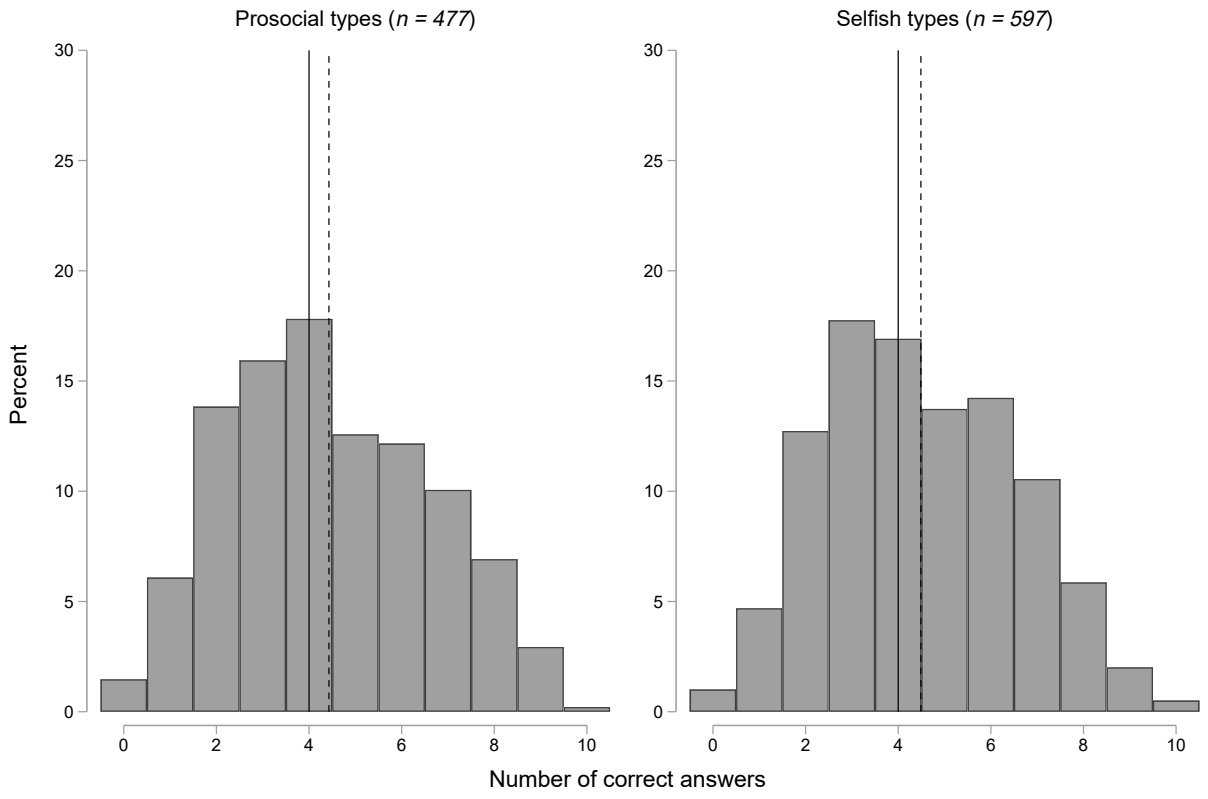


Figure A7: Distribution of number of correct answers in the IQ task. Dashed lines display the mean score, solid lines display the median.

Table A1: Overview of experiments

| Experiment | Lab | # Participants | $\delta$ | Matching | Group composition announced | # Supergames | # Groups (prosocial, mixed, selfish) | % prosocial types | Avg. payoff |
|---|---|---|---|---|---|---|---|---|---|
| Main experiment | Cologne | $n = 240$ | 0.6 | sorted | yes | 20 | 7, 6, 11 | 41.7 | 17.36€ |
| Random matching | Cologne | $n = 78$ | 0.6 | random | no | 20 | -, 9, - | 39.7 | 16.31€ |
| Unannounced matching. | Bonn | $n = 228$ | 0.6 | sorted | no | 20 | 7, 8, 9 | 44.4 | 16.98€ |
| High delta | Cologne | $n = 270$ | 0.8 | sorted | yes | 20 | 8, 8, 11 | 42.6 | 17.88€ |
| Long-run | Bonn | $n = 198$ | 0.6 | sorted | yes | 40 | 8, 8, 8 | 52.0 | 17.05€ |

Table A2: 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 | 43.48 | 1.68 | 16.20 | 1.49 |
| Cooperate | 9.50 | 0.93 | 24.12 | 2.61 |
| Total | 52.98 | 2.61 | 40.32 | 4.10 |

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

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: Cooperation rates across supergames and group type ($\delta = 0.6$, unannounced matching)

| | First round | | | All rounds | | |
|---|---|---|---|---|---|---|
| | 1 | 11-20 | All | 1 | 11-20 | All |
| Prosocial groups | 0.64 | 0.67 | 0.68 | 0.60 | 0.52 | 0.56 |
| Selfish groups | 0.26 | 0.17 | 0.22 | 0.17 | 0.09 | 0.14 |
| Mixed groups | 0.41 | 0.42 | 0.46 | 0.31 | 0.29 | 0.34 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p = 0.003$ | $p = 0.014$ | $p = 0.002$ | $p = 0.009$ | $p = 0.040$ | $p = 0.025$ |
| $H_0$: Mixed = Selfish | $p = 0.062$ | $p = 0.002$ | $p = 0.001$ | $p = 0.008$ | $p = 0.001$ | $p < 0.001$ |

*Notes:* Differences between group types are tested using probit regressions with standard errors clustered at the matching group level. Very similar results are obtained when only looking at last round cooperation (see Table C2 in Appendix C).

Table A5: Cooperation rates across supergames and group type ($\delta = 0.8$)

| | First round | | | All rounds | | |
|---|---|---|---|---|---|---|
| | 1 | 11-20 | All | 1 | 11-20 | All |
| Prosocial groups | 0.77 | 0.83 | 0.85 | 0.50 | 0.76 | 0.75 |
| Selfish groups | 0.26 | 0.37 | 0.40 | 0.15 | 0.22 | 0.26 |
| Mixed groups | 0.56 | 0.64 | 0.67 | 0.47 | 0.55 | 0.61 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p = 0.004$ | $p < 0.001$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p < 0.006$ | $p = 0.103$ | $p = 0.079$ | $p = 0.106$ | $p < 0.001$ | $p = 0.187$ |
| $H_0$: Mixed = Selfish | $p < 0.001$ | $p = 0.025$ | $p = 0.013$ | $p = 0.004$ | $p = 0.008$ | $p < 0.001$ |

*Notes:* Differences between group types are tested using probit regressions with standard errors clustered at the matching group level. Very similar results are obtained when only looking at last round cooperation (see Table C3 in Appendix C).

Table A6: Long-run cooperation rates across supergames and group type ($\delta = 0.6$)

| | First round | | | All rounds | | |
|---|---|---|---|---|---|---|
| | 1-20 | 21-40 | All | 1-20 | 21-40 | All |
| Prosocial groups | 0.67 | 0.61 | 0.64 | 0.52 | 0.49 | 0.50 |
| Selfish groups | 0.31 | 0.22 | 0.26 | 0.23 | 0.16 | 0.20 |
| Mixed groups | 0.48 | 0.43 | 0.45 | 0.40 | 0.36 | 0.38 |
| $H_0$: Prosocial = Selfish | $p = 0.002$ | $p = 0.009$ | $p = 0.003$ | $p = 0.002$ | $p = 0.015$ | $p = 0.006$ |
| $H_0$: Prosocial = Mixed | $p = 0.201$ | $p = 0.276$ | $p = 0.232$ | $p = 0.350$ | $p = 0.435$ | $p = 0.392$ |
| $H_0$: Mixed = Selfish | $p = 0.156$ | $p = 0.095$ | $p = 0.115$ | $p = 0.073$ | $p = 0.080$ | $p = 0.072$ |

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

Table A7: Estimated strategy frequencies ($\delta = 0.6$, unannounced matching)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | $0.234^{***}$ | $0.422^{***}$ | $0.804^{***}$ |
| | (0.041) | (0.047) | (0.041) |
| Always cooperate (AC) | $0.052^{**}$ | $0.023^{*}$ | 0.000 |
| | (0.026) | (0.013) | (0.000) |
| Grim trigger (GT) | $0.320^{***}$ | $0.328^{***}$ | $0.082^{***}$ |
| | (0.037) | (0.052) | (0.027) |
| Tit-for-tat (TFT) | $0.345^{***}$ | $0.160^{***}$ | $0.060^{***}$ |
| | (0.041) | (0.034) | (0.023) |
| Suspicious Tit-for-tat (STFT) | $0.048^{**}$ | $0.067^{**}$ | $0.054^{**}$ |
| | (0.020) | (0.026) | (0.027) |
| Gamma | $0.455^{***}$ | $0.507^{***}$ | $0.447^{***}$ |
| | (0.022) | (0.026) | (0.022) |
| Frequency of cooperative strategies | 0.766 | 0.578 | 0.196 |
| Observations | 88 | 86 | 114 |

*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. See Tables C6 and C7 in Appendix C for robustness checks.

Table A8: Estimated strategy frequencies ($\delta = 0.8$)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.100** | 0.215*** | 0.491*** |
| | (0.041) | (0.066) | (0.067) |
| Always cooperate (AC) | 0.141*** | 0.213*** | 0.000 |
| | (0.051) | (0.051) | (0.002) |
| Grim trigger (GT) | 0.243*** | 0.179*** | 0.062*** |
| | (0.067) | (0.048) | (0.021) |
| Tit-for-tat (TFT) | 0.515*** | 0.322*** | 0.375*** |
| | (0.072) | (0.060) | (0.068) |
| Suspicious Tit-for-tat (STFT) | 0.000 | 0.071*** | 0.072*** |
| | (0.000) | (0.033) | (0.025) |
| Gamma | 0.401*** | 0.389*** | 0.446*** |
| | (0.033) | (0.019) | (0.028) |
| Frequency of cooperative strategies | 0.900 | 0.785 | 0.509 |
| Observations | 80 | 80 | 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. See Table C8 in Appendix C for a robustness check.

Table A9: Predictors of decision to cooperate conditional on partner's cooperation

| | (1) |
|---|---|
| Female | 0.017 |
| | (0.034) |
| Norm following | 0.171*** |
| | (0.001) |
| Raven test score | 0.010 |
| | (0.007) |
| General risk attitude | -0.033 |
| | (0.007) |
| Agreeableness | 0.079*** |
| | (0.015) |
| Conscientiousness | -0.086*** |
| | (0.015) |
| Extraversion | 0.010 |
| | (0.012) |
| Neuroticism | 0.023 |
| | (0.012) |
| Openness | 0.084** |
| | (0.012) |
| Observations | 1074 |

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

# 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 $r$.[11] 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 $\theta^k$ and the variance $\gamma$ of the error term $\varepsilon_{i,g,r}$, among individuals in a given

[11] 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 Robustness Checks

Table C1: Last round cooperation rates across supergames and group type ($\delta = 0.6$)

| | Supergame | | |
|---|---|---|---|
| | 1 | 11-20 | All |
| Prosocial groups | 0.74 | 0.54 | 0.60 |
| Selfish groups | 0.18 | 0.09 | 0.13 |
| Mixed groups | 0.33 | 0.23 | 0.26 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p < 0.001$ | $p = 0.053$ | $p = 0.019$ |
| $H_0$: Mixed = Selfish | $p = 0.005$ | $p = 0.256$ | $p = 0.229$ |

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

Table C2: Last round cooperation rates across supergames and group type ($\delta = 0.6$, unannounced matching)

| | Supergame | | |
|---|---|---|---|
| | 1 | 11-20 | All |
| Prosocial groups | 0.55 | 0.54 | 0.58 |
| Selfish groups | 0.11 | 0.09 | 0.13 |
| Mixed groups | 0.30 | 0.30 | 0.35 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p = 0.005$ | $p = 0.020$ | $p = 0.014$ |
| $H_0$: Mixed = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |

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

Table C3: Last round cooperation rates across supergames and group type ($\delta = 0.8$)

| | Supergame | | |
|---|---|---|---|
| | 1 | 11-20 | All |
| Prosocial groups | 0.59 | 0.76 | 0.74 |
| Selfish groups | 0.11 | 0.26 | 0.27 |
| Mixed groups | 0.43 | 0.56 | 0.59 |
| $H_0$: Prosocial = Selfish | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| $H_0$: Prosocial = Mixed | $p = 0.155$ | $p = 0.120$ | $p = 0.169$ |
| $H_0$: Mixed = Selfish | $p < 0.001$ | $p = 0.015$ | $p = 0.003$ |

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

Table C4: Estimated strategy frequencies excluding supergames with only one round ($\delta = 0.6$)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.268*** | 0.662*** | 0.717*** |
| | (0.102) | (0.133) | (0.068) |
| Always cooperate (AC) | 0.110** | 0.034 | 0.003 |
| | (0.048) | (0.038) | (0.021) |
| Grim trigger (GT) | 0.193** | 0.106* | 0.041 |
| | (0.082) | (0.054) | (0.041) |
| Tit-for-tat (TFT) | 0.408*** | 0.164** | 0.067** |
| | (0.082) | (0.076) | (0.034) |
| Suspicious Tit-for-tat (STFT) | 0.021 | 0.034 | 0.172** |
| | (0.022) | (0.033) | (0.068) |
| Gamma | 0.482*** | 0.449*** | 0.412*** |
| | (0.045) | (0.048) | (0.050) |
| Frequency of cooperative strategies | 0.732 | 0.338 | 0.283 |
| Observations | 70 | 60 | 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 C5: Estimated strategy frequencies excluding the first 10 supergames ($\delta = 0.6$)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.306*** | 0.777*** | 0.674*** |
| | (0.096) | (0.150) | (0.159) |
| Always cooperate (AC) | 0.062 | 0.052 | 0.000 |
| | (0.051) | (0.043) | (0.007) |
| Grim trigger (GT) | 0.206** | 0.093 | 0.079 |
| | (0.081) | (0.078) | (0.060) |
| Tit-for-tat (TFT) | 0.396*** | 0.079 | 0.000 |
| | (0.115) | (0.071) | (0.005) |
| Suspicious Tit-for-tat (STFT) | 0.031 | 0.000 | 0.246 |
| | (0.043) | (0.039) | (0.189) |
| Gamma | 0.436*** | 0.402*** | 0.349*** |
| | (0.058) | (0.052) | (0.058) |
| Frequency of cooperative strategies | 0.694 | 0.223 | 0.326 |
| Observations | 70 | 60 | 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 C6: Estimated strategy frequencies excluding supergames with only one round ($\delta = 0.6$, unannounced matching)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.272*** | 0.423*** | 0.793*** |
| | (0.042) | (0.046) | (0.041) |
| Always cooperate (AC) | 0.054** | 0.023* | 0.000 |
| | (0.026) | (0.013) | (0.000) |
| Grim trigger (GT) | 0.273*** | 0.316*** | 0.090*** |
| | (0.040) | (0.052) | (0.026) |
| Tit-for-tat (TFT) | 0.343*** | 0.131*** | 0.056*** |
| | (0.043) | (0.030) | (0.020) |
| Suspicious Tit-for-tat (STFT) | 0.057** | 0.107*** | 0.061** |
| | (0.023) | (0.031) | (0.026) |
| Gamma | 0.452*** | 0.472*** | 0.413*** |
| | (0.021) | (0.024) | (0.020) |
| Frequency of cooperative strategies | 0.728 | 0.577 | 0.207 |
| Observations | 88 | 86 | 114 |

*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 C7: Estimated strategy frequencies excluding the first 10 supergames ($\delta = 0.6$, unannounced matching)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.261*** | 0.473*** | 0.810*** |
| | (0.043) | (0.046) | (0.042) |
| Always cooperate (AC) | 0.073** | 0.037** | 0.000 |
| | (0.030) | (0.017) | 0.000 |
| Grim trigger (GT) | 0.339*** | 0.208*** | 0.057** |
| | (0.051) | (0.042) | (0.024) |
| Tit-for-tat (TFT) | 0.287*** | 0.180*** | 0.061** |
| | (0.046) | (0.040) | (0.026) |
| Suspicious Tit-for-tat (STFT) | 0.039** | 0.102*** | 0.072** |
| | (0.018) | (0.033) | (0.032) |
| Gamma | 0.391*** | 0.404*** | 0.367*** |
| | (0.019) | (0.021) | (0.021) |
| Frequency of cooperative strategies | 0.739 | 0.427 | 0.190 |
| Observations | 88 | 86 | 114 |

*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 C8: Estimated strategy frequencies excluding the first 10 supergames ($\delta = 0.8$)

| | Prosocial groups | Mixed groups | Selfish groups |
|---|---|---|---|
| Always defect (AD) | 0.125*** | 0.229*** | 0.497*** |
| | (0.044) | (0.073) | (0.066) |
| Always cooperate (AC) | 0.156*** | 0.227*** | 0.027 |
| | (0.059) | (0.058) | (0.020) |
| Grim trigger (GT) | 0.228*** | 0.191*** | 0.085** |
| | (0.066) | (0.057) | (0.040) |
| Tit-for-tat (TFT) | 0.480*** | 0.241*** | 0.276*** |
| | (0.073) | (0.058) | (0.064) |
| Suspicious Tit-for-tat (STFT) | 0.011 | 0.112* | 0.115*** |
| | (0.011) | (0.057) | (0.033) |
| Gamma | 0.344*** | 0.360*** | 0.385*** |
| | (0.032) | (0.030) | (0.028) |
| Frequency of cooperative strategies | 0.875 | 0.771 | 0.503 |
| Observations | 80 | 80 | 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.

# D Experimental Instructions

*Shown are the instructions for the main experiment (translated from German).*

## 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.

## 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

Assume that 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?

## 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) chose to play A in case the other player chooses A.[12] all participants in your group (including you) chose to play B in case the other player chooses A.[13] some participants in your group chose A in case the other player choose A and some chooe B in case the other player chooses A.[14]

### 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.

[12] Displayed to participants in prosocial groups.
[13] Displayed to participants in selfish groups
[14] Displayed to participants in 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?

# E 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.

## 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".


Figure E1: Norm Following Decision screen norm-following task