Cooperation in infinitely repeated Prisoner Dilemmas: unexplained variation and social preferences

Michela Boldrini *

October 2020

Contents

* Department of Economics, University of Bologna: P.zza Scaravilli 2, 40126 - Bologna (Italy)
E-mail address: michela.boldrini2@unibo.it
We thank all the authors who made the data we use in this article available.

1 Introduction

In this project we aim to investigate the determinants of individuals' cooperative behavior in infinitely repeated Prisoner Dilemmas (henceforth PDs).

Our main contribution is to shed light on what is the role of individuals' social preferences in shaping cooperation across contexts that differ in terms of strategic incentives for cooperation, which are defined by the environmental game parameters. In absence of conclusive evidence from the previous experimental literature, our approach allows us to neatly test whether social preferences have a *per se* effect on cooperation and to analyze whether the effect of social preferences is confounded in contexts where cooperation could be sustainable even in absence of social preferences. The ultimate objective is therefore to bridge the two strands of the experimental literature that, separately, looked at how strategic incentives, on one side, and individual preferences and characteristics, on the other side, can account for subjects' cooperativeness in infinitely repeated PDs (see Roth and Murnighan

[1978],Murnighan and Roth [1983], Dal Bó [2005], Blonski et al. [2011],Dal Bó and Fréchette [2011], Dal Bó and Fréchette [2018] for the former strand of the literature and Sabater-Grande and Georgantzis. [2002], Dreber et al. [2014], Davis et al. [2016], Proto et al. [2019] for the latter), in the attempt to better interpret some of the so-far unexplained variation in cooperation levels observed in experimental data.

First, we present a meta-analysis run on an extended version of the dataset collected by Dal Bó and Fréchette [2018], where we rely on simple supervised-learning algorithms to test the ability of environmental game parameters to predict cooperation. This analysis confirms that environmental game parameters (such as payoffs and continuation probabilities) do have some predictive power, especially with respect to choices taken after that subjects have gained some experience. In particular, the two composite indicators derived by Dal Bó and Fréchette [2011] ($sizeBAD$) and Blonski et al. [2011] ($\delta - \delta_{RD}$) to explain cooperation levels seem to, indeed, capture a great share of the information relevant to predict cooperation but their predictive accuracy is steadily lower in contexts where cooperation is not sustainable as a long-run equilibrium, based on environmental game parameters.

Later, we present the theoretical framework through which we model how social preferences could affect cooperativeness and the experimental approach we rely on to empirically investigate the role of social preferences through an online experiment.
Our experimental design allows us to separately "measure" subjects' social preferences - through an experimental procedure inspired by the approach developed by Bruhin et al. [2018] - and observe subjects' actual cooperativeness in One Shot and Infinitely Repeated PDs.
The data collected through our experimental procedure allow us to test: (1) whether and how social preferences affect subjects' cooperativeness in One Shot and Infinitely Repeated PDs, (2) how the degree of concentration of subjects who exhibit strong social preferences can affect the levels of cooperation attained in in One Shot and Infinitely Repeated PDs, and whether and how (3) social preferences affect subjects' sensitivity to strategic incentives for ooperation, set by environmental game parameters.

From the meta-analysis, we find that strategic incentives set by game parameters predict cooperation increasingly better over supergames and asymmetrically in treatments where cooperation is and is not sustainable as an equilibrium.

The experimental evidence, instead, proves that social preferences do play a relevant role in shaping cooperative attitudes both in One Shot and Infinitely Repeated PDs. It also highlights the key role played by beliefs, which explain a large fraction of the variation observed, serving as a transmission channel for the effect of social preferences on behavior.

2 Motivation & Literature Review

The study of social dilemmas, such as PDs, where individual and collective interests are in conflict, has long attracted the interest of economists, who contributed to this field of research both theoretically and experimentally. In particular, studying how individuals behave in contexts where they face the same social dilemma an indefinite number of times, is extremely relevant from a policy perspective since these contexts more closely mirror real-world situations where subjects are not informed ex-ante of the duration of their future interactions with others.

Contrary to the case of One Shot or finitely-repeated interactions, however, when we introduce infinitely repeated interactions, standard economic theory fails to postulate univocal predictions on subjects' behavior, opening for the possibility that even among purely self-interested individuals cooperation could be sustained in equilibrium, if players are sufficiently patient [1].

Likewise, the empirical evidence collected so far did not succeed in isolating what factors can best predict the emergence of cooperative long-run equilibria and in explaining the heterogeneity in cooperativeness observed in contexts that should be equivalent to subjects from a theoretical point of view.

In a canonical 2x2 PD, see Table 1, subjects face a binary choice on whether to cooperate or defect, given the following payoffs:

T: Temptation's payoff from defecting when the other cooperates

[1] Folk Theorem, see Fudenberg and Maskin [1986]

R: Reward from mutual cooperation

P: Punishment from mutual defection

S: Sucker's payoff from cooperating when the other defects

where $T > R > P > S$ and, typically, $2R > (T + S)$, which makes joint cooperation more profitable than alternating between cooperation and defection.

Table 1: Prisoners' Dilemma Row Player's Payoffs

Original

	C	D
C	R	S
D	T	P

Normalized

	C	D
C	$\frac{R-P}{R-P} = 1$	$\frac{S-P}{R-P} = -l$
D	$\frac{T-P}{R-P} = 1 + g$	$\frac{P-P}{R-P} = 0$

If we consider the normalized version [2] of the payoffs' matrix, the number of relevant parameters in the stage game reduces to two: the gain from unilateral defection (g) and the loss from unilateral cooperation (l). When they are implemented in the laboratory, following the pioneering contribution of Roth and Murnighan [1978] and Murnighan and Roth [1983], infinitely repeated PDs essentially transform into 'indefinitely repeated' games where subjects play the stage PD game an indefinite number of times and new relevant parameters emerge: in 'supergame', subjects are matched to a partner and play the stage PD game with the same partner a number of rounds that depends on a pre-set 'continuation probability' (δ); when the supergame is over, subjects are re-matched to new partners and play the same repeated PD game again; this procedure is iterated for each supergame, with the total number of supergames to be played being pre-determined.

Two strands of the experimental literature on infinitely repeated PDs, so far largely unrelated,

[2] The normalized version of the payoffs' matrix is obtained by applying a monotonic linear transformation to the original matrix.

focused on the role of environmental game parameters (like the continuation probability δ, the gain from unilateral defection g, etc.) on cooperation, on one side, and on the role of individual preferences and characteristics, on the other side.

The strand of the economic literature focusing on the role of environmental game parameters leveraged on the most recent advances in the theory of infinitely repeated games to study whether and to what extent these parameters can affect cooperation levels in PDs. A recent work by Dal Bó and Fréchette [2018] offers a comprehensive review of the main contributions on this topic (Roth and Murnighan [1978],Murnighan and Roth [1983], Dal Bó [2005], Blonski et al. [2011],Dal Bó and Fréchette [2011]) while providing some empirical evidence on how environmental game parameters affect cooperation by relying on a set of meta-data that brings together more than 150.000 observations collected from 15 different experimental papers. Cooperation is generally found to be increasing in the probability of future interactions and, on average, greater when cooperation can be supported as a Subgame Perfect Nash Equilibrium (SPE) or as a Risk Dominant Equilibrium (RD) [3] , although a large amount of variation is left unexplained. In the attempt to dig deeper in the unexplained variation in cooperation levels observed, different approaches have been followed, which tried to best combine and compact the information contained by envirmental game parameters into composite indicators: the two most prominent examples are provided by Blonski et al. [2011], who build a continuous measure of 'how risk-dominant' is cooperation [4] based on the distance between δ and δ_{RD} (where $\delta_{RD} = \frac{g+l}{1+g+l}$), and by Dal Bó and Fréchette [2011], who build a continuous measure of how resistant is cooperation to strategic uncertainty based on the basin of attraction of the Always Defect (AD) strategy ($sizeBAD$) [5]. Dal Bó and Fréchette [2018] separately test the ability of these two composite indicators to predict cooperation in their meta-data: they show

[3]The concept of Risk Dominance is borrowed from the literature on coordination games, where Harsanyi et al. [1988] define an equilibrium to be risk-dominant to another if the opportunity costs of unilaterally deviating from that equilibrium is higher. Blonski et al. [2011] and Blonski and Spagnolo [2015] develop an equilibrium selection theory for infinitely repeated PDs, moving from the concept of 'strategic risk' by Harsanyi et al. [1988].

[4]Assuming subjects are uncertain about their opponent's moves, we consider a strategy to be *risk-dominant* if it is a best-response to the other player randomizing with a 50-50% probability between a cooperative strategy (Grim) and a non-cooperative strategy (Always Defect).

[5]The basin of attraction of AD against a cooperative strategy like Grim corresponds to the maximum probability of the other player playing Grim that makes playing AD optimal. When cooperation can be supported in equilibrium it is equal to $\frac{(1-\delta)l}{(1-(1-\delta)(1+g-l))}$. When cooperation is not supported in equilibrium this maximum probability is equal to 1.

that cooperation is positively correlated with the distance $(\delta - \delta_{RD})$, especially when cooperation is Risk Dominant (treatments where $\delta > \delta_{RD}$), and negatively correlated to the size of the basin of attraction of AD ($sizeBAD$) when cooperation is Risk Dominant. The question on which of the two indices predicts best cooperation is left unanswered, given the high correlation between the two indices in the meta-data.

Another growing strand of the literature, instead, recently focused on the role of individual characteristics - including preferences - on cooperation in infinitely repeated games. The role of individual preferences on cooperation has already been analyzed in the context of finitely repeated games, where free-riding and defection are the only possible outcomes for rational and self interested individuals, and thus preferences - in particular social preferences - are deemed necessary to justify the emergence of cooperation (Fehr and Fischbacher [2003]). Some studies focused on the role of social preferences on cooperation in One Shot PDs and found some evidence of a positive correlation between cooperation and other-regarding attitudes, which were typically measured through other games: Blanco et al. [2011] report the presence of a positive correlation between cooperative behavior in a sequential One Shot PD and other-regarding behavior measured in terms of giving in a modified dictator game and in an ultimatum game, and likewise, Capraro et al. [2014] find a positive correlation between cooperative behavior in a One Shot continuous-choice PD and giving in a dictator game. These results seem to support the hypothesis by Peysakhovich et al. [2014] that individuals display a "cooperative phenotype", which is not correlated with norm-enforcing punishment or non-competitiveness but is valid in a general domain and substantiates in a temporally stable inclination towards paying costs to benefit others that makes subjects' behavior consistent across different decision scenarios.

The picture appears, however, different when we introduce infinite repetition. Dal Bó and Fréchette [2018] offer a comprehensive review of the main findings from this strand of the literature, where no conclusive evidence has yet been found in favor of the presence of a systematic effect of individual characteristics such as risk aversion (Sabater-Grande and Georgantzis. [2002], Dreber et al. [2014], Davis et al. [2016], Proto et al. [2019]), social preferences (Dreber et al. [2014], Davis et al. [2016]), intelligence (Proto et al. [2019]), patience (Davis et al. [2016], Kim [2019]) etc. on cooperative-

ness. Interestingly, the effect of some of these characteristics seems to be sensitive to the strategic environment defined by the parameters of the infinitely repeated game, as it is the case for social preferences, which are found to be predictive of cooperation only when cooperation is not sustainable as an equilibrium: Dreber et al. [2014], by letting subjects play a standard dictator game (DG) after a series of indefinitely repeated PDs where cooperation either is or is not an equilibrium (between-subjects design), find that cooperation is related to giving in the DG only when the parameters of the PD game make cooperation not sustainable as a long-run equilibrium. Similarly, Arechar et al. [2018], who let their subjects play a standard DG before and after an infinitely repeated PD with varying continuation probability (within-subjects design), find that the giving behavior in the first DG predicts both the level of cooperation and the strategies played by subjects in the PD: givers are found more likely to cooperate and less likely to choose a non-cooperative strategy like 'Always Defect", but only when cooperation is not sustainable as an equilibrium. Consistently, Davis et al. [2016], who let their subjects play an infinitely repeated PD where cooperation is an equilibrium and later ask the same subjects to perform a series of tasks aimed to measure their personal attitude over a series of dimensions, find that cooperation is not systematically related to social preferences, measured in terms of altruism and behavior in a trust game.

This evidence could be consistent with the idea that individuals could also have a *non-strategic* taste for cooperation, in addition to a *strategic* taste for cooperation. In this framework, which we will further describe in the dedicated section, individuals having a strategic taste for cooperation would exhibit a cooperative attitude only when the environmental characteristics of the game and their expectations are such that cooperation culd be a profitable strategy from a purely rational point of view, while individuals having a non-strategic taste for cooperation would exhibit a cooperative attitude even when the environmental characteristics of the game and their expectations do not guarantee cooperation to be a profitable strategy. If this was the case, environmental characteristics of the game could be effective predictors of cooperativeness when cooperation is sustainable as an equilibrium but not necessarily otherwise, while, vice-versa, factors explaining the non-strategic taste for cooperation could be relevant predictors of cooperativeness when cooperation is not an equilibrium but not necessarily otherwise.

3 Meta-Analysis

We perform a meta-analysis on a wide set of data collected from previous experimental works on Infinitely Repeated PDs, relying on some simple off-the-shelf (Athey [2017]) supervised learning algorithms, in order to test the ability of enviornmental game parameters to predict cooperation.
We rely on machine learning (ML) techniques because we are interested in testing the "empirical" relevance of environemental game parameters without imposing too much structure, with the objective to uncover regularities that could also be generalized to other PD settings that are not included in our sample. In this context, ML techniques serve well our purposes, allowing for a high degree of flexibility in the model structure, which is derived through a purely a-theoretical and data-driven procedure aimed to maximize the out-of-sample prediction performance of the model. [6].
Although ML applications on experimental data have only recently started to grow (Fudenberg and Peysakhovich [2016]),Naecker [2015],Nay and Vorobeychik [2016], Naecker and Peysakhovich [2017], Fudenberg and Liang [2019]) it emerged that ML algorithms can serve as useful instruments for experimental economists, allowing them to uncover unexplored regularities in the data that can help in better explaining the mechanisms behind subjects' behavior, complementing the information provided by theoretical models.
Through this meta-analysis we will 'let the data speak' on the relative predictive performace of the different environmental game parameters considered by the theory, in order to test: (i) how well these parameters can predict Round-1 cooperation choices; (ii) whether the composite indicators derived from the theory - $(\delta - \delta_{RD})$ and $sizeBAD$ - would be selected among the most relevant predictors by completely a-theoretical routines solely based on parameters' ability to describe the patterns observed in the data.

[6] ML techniques rely on highly flexible functional forms, where greater complexity improves the in-sample fit but increases the out-of-sample error of the selected model: the level of complexity of the model is then set through a regularization parameter that is chosen by cross validation in order to minimize the out-of-sample error (Hastie et al. [2009]).

The analysis is conducted over an extended version of the dataset collected by Dal Bó and Fréchette [2018], which brings together data from 15 different randomly-terminated 'standard' PD experiments with perfect monitoring and fixed matching across supergames [7].

We collect data from 18 papers, involving 33 different treatments - identified as non-overlapping combinations of environmental parameters δ, l, g - with a total of 3267 subjects and observations on almost 270.000 choices.

We focus on Round-1 cooperation choices across different supergames, although it may be argued that they only provide an imcomplete picture of individuals' cooperative attitude, because it simplifies our analysis over a series of dimensions: first, different treatments and different supergames within the same treatment may have a different number of rounds, which would complicate the analysis; second, Round-1 choices can be though as solely reflecting subjects' own individual strategies, net of the impact of other players' strategies that are likely to impact subjects' subsequent choices in the supergame; third the binary choice of cooperating/defecting in the first round can be univocally mapped into subjects' willingness to engage in a cooperative (Always Cooperate, Grim, etc.) or non-cooperative (Always Defect) strategy, without imposing restrictions on the set of possible strategies.

Our main focus will be on supergames 1 to 7, as the 7th supergame is the highest supergame we can study without losing any treatment. We run the analysis separately for each supergame. The algorithm is fed with a series of treatment-specific characteristics of the game (S_t) that include:

[7] Dal Bó and Fréchette [2018] main inclusion criteria are: (1) the stage game is a fixed 2x2 PD game; (2) there is perfect monitoring; (3) there's One Shot interaction or repeated interaction through a random continuation rule (and this does not change inside a supergame); (4) pairs are fixed inside a supergame. They further condition the inclusion on the data availability and on a publication date no before 2014, if the paper is not their own.
We follow the same inclusion criteria extending the publication date up until 2019 and conditioning on the availability of at least 7 supergames per treatment. We restrict our attention to papers where the authors report detailed information on subjects' payments. We use Internet searches to find articles satisfying these conditions. The papers included in our meta-analysis are: Andreoni and Miller. [1993]; Cooper et al. [1996], Dal Bó [2005], Dreber et al. [2008], Aoyagi and Freéchette. [2009], Duffy and Ochs [2009], Dal Bó et al. [2010], Dal Bó and Fréchette [2011], Blonski et al. [2011], Fudenberg et al. [2012], Bruttel and Kamecke [2012], KaterinaSherstyuk et al. [2013], Fréchette and Yuksel [2017], Dal Bó and Fréchette [2019], Peysakhovich and Rand [2016], Romero and Rosokha [2018], Ghidoni et al. [2019] and Proto et al. [2019]
In the Appendix (Section A.1) we replicate one of the main analysis proposed by Dal Bó and Fréchette [2018] to study the effect of $\delta - \delta_{RD}$ and $sizeBAD$ on cooperation on our dataset, in order to show that it is not systematically different from the dataset originally collected by Dal Bó and Fréchette [2018].
Table A1.3 provides details on the variety of treatments encompassed in our dataset, including those that are also included in the analysis run by Dal Bó and Fréchette [2018].

- δ: the continuation probability
- g: the 'normalized' gain from unilateral defection
- l: the 'normalized' loss from unilateral cooperation
- I_{RD}: a dummy for treatments where cooperation is a Risk Dominant Eq. (RD)
- I_{SPE}: a dummy for treatments where cooperation is a Subgame Perfect Nash Eq. (SPE)
- $matching$: a variable describing the between-supergames matching protocol [8]
- $m_{(R-P)}$: the mark-up from mutual cooperation with respect to mutual defection that is equal to the distance between R and P payoffs from the original matrix [9]
- $P_{DDpayoff}$: the payoff from mutual defection from the original matrix [10]
- $totsup$: the total number of supergames
- $showup$: the amount of the show-up fee [11]

In order to evaluate the predictive accuracy of the two indicators $(\delta - \delta_{RD})$ and $sizeBAD$, we separately estimate the predictive model for each supergame three times, employing slightly different input sets, see Table 2: first, we estimate the model including all treatment characteristics and the $(\delta - \delta_{RD})$ indicator only $(S_t^1 = S_t + (\delta - \delta_{RD}))$; second, we estimate the model including all treatment characteristics and the $sizeBAD$ indicator only $(S_t^2 = S_t + sizeBAD)$; third, we estimate the model including all treatment characteristics and both the $(\delta - \delta_{RD})$ and $sizeBAD$ indicators $(S_t^3 = S_t + (\delta - \delta_{RD}) + sizeBAD)$.

Table 3 shows the list of non-null coefficients selected by the Logistic LASSO algorithm throughout supergames [12] : composite indicators $(\delta - \delta_{RD})$ and $sizeBAD$ are always selected among the most

[8] $matching$ is a ordered discrete variable that takes value '1' if subjects were matched according to complete stranger protocol, aka turnpike or zipper; takes value '2' if subjects were matched according to perfect stranger protocol, aka round robin; takes value '3' if subjects were matched according to perfect stranger protocol until the pool is exhausted; takes value '4' if subjects were matched according to a random matching protocol.

[9] The amount is measured in US dollars: if the experiments were originally paid in dollars, the amounts are obtained by multiplying experimental currency units (ECU) by the exchange rate to dollars; if the experiments were originally paid in other currencies, the amounts are obtained by (1) multiplying experimental currency units (ECU) by the exchange rate to the relevant currency (2) converting the amounts obtained to american dollars. The same procedure was adopted by Dal Bó and Fréchette [2018].
Blonski et al. [2011] was conducted between May and November of 2006; accordingly the exchange rate is set at 0.785 Euro = 1 dollar. Bruttel and Kamecke [2012] was conducted between January and February of 2007; accordingly the exchange rate is set at 1 Euro = 1.307 dollars. Ghidoni et al. [2019] was conducted between June 2016 and November 2017; accordingly the exchange rate is set at 1 Euro = 1.1156 dollars. Proto et al. [2019] was conducted between June 2013 and June 2016; accordingly the exchange rate is set at 1 GBP = 1.5718 dollars.

[10] see footnote 9.

[11] see footnote 9

[12] We employ three different supervised learning algorithms to deal with our classification problem, where $Y = 1[Round - 1choice = cooperation]$: the Decision Tree, the Random Forest and the Logistic LASSO.
Before fitting our models we randomly split our sample in two subsamples, a training set and a testing set, following

relevant cooperation predictors, either alone or in combination with other game parameters.

Tables 4 and 5 report some classification accuracy metrics for Round-1 cooperation choices in Supergame 1 and 7, respectively. [13].

Table 2: Logistic LASSO: set of inputs

(1) $\delta - \delta_{RD}$ **model**	**(2)** $sizeBAD$ **model**	**(3) Unconstrained model**
S_t^1: treatment characteristics $\delta,\ g,\ l,\ I_{RD},\ I_{SPE},$ $matching,\ m_{(R-P)},$ $P_{DDpayoff},\ showup,$ $totsup,\ \delta - \delta_{RD}$	S_t^2: treatment characteristics $\delta,\ g,\ l,\ I_{RD},\ I_{SPE},$ $matching,\ m_{(R-P)},$ $P_{DDpayoff},\ showup,$ $totsup,\ sizeBAD$	S_t^3: treatment characteristics $\delta,\ g,\ l,\ I_{RD},\ I_{SPE},$ $matching,\ m_{(R-P)},$ $P_{DDpayoff},\ showup,$ $totsup,\ \delta - \delta_{RD}\ ,\ sizeBAD$
I_t^1: all pairwise interactions between variables in S_t^1	I_t^2: all pairwise interactions between variables in S_t^2	I_t^3: all pairwise interactions between variables in S_t^3
$\tilde{X}_t^1$ contains 66 predictors	$\tilde{X}_t^2$ contains 66 predictors	$\tilde{X}_t^3$ contains 78 predictors

the 2/3 and 1/3 division rule (as suggested by Y. Zhao and Cen [2014]). We then fit our models on the training set and test the predictive accuracy of the models over the testing set. We chose to report only the results obtained by using the Logistic LASSO algorithm as it shows a good predictive performance with respect to the other algorithms in most of the cases and, at the same time, it performs well on the ground of interpretability, providing easy-to-interpret outputs. See the Appendix (Section A.1) for more details on the Logistic LASSO algorithm.

[13]The logistic LASSO outputs a predicted probability to cooperate for each individual in the sample. However, since we feed the algorithm with treatment-specific variables only, without including any individual-specific information, we obtain the same estimated probability to cooperate for all individuals exposed to the same treatment.
Standard metrics used to evaluate predition accuracy of ML classification algorithms typically convert predicted probabilities into predicted categories by setting a predicted probability cutoff c such that those observations whose predicted probability is above c are classified as belonging to the $Y = 1$ category, while the others are classified as belonging to the $Y = 0$ category. In this context, given that the predicted probability is the same for all individuals exposed to the same treatment, we would have that all individuals exposed to the same treatment would either be classified as Round-1 cooperators or defectors with no within-treatment variability: misclassified individuals will then be cooperators exposed to treatments where the predicted probability to cooperate is below the cutoff or defectors exposed to treatments where the predicted probability to cooperate is above the cutoff.
The area under the ROC (Receiver Operating Characteristics) Curve provides a measure of classification accuracy considering the entire set of all possible thresholds $c \in [0, 1]$, while the other metrics assume the predicted probability cutoff for classification is set at $c = 0.5$.

Table 3: Logistic LASSO: non-null estimated coefficients - Supergame 1 to 7

(1) $\delta - \delta_{RD}$ **model**	**(2)** $sizeBAD$ **model**	**(3) Unconstrained model**
	Supergame 1	
(+) $I_{RD}\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$ $m \cdot totsup$, $I_{SPE} \cdot totsup$ (-) $l, g \cdot l$	(+) $I_{RD} \cdot totsup$ (-) $sizeBAD$, $l, g\cdot\ sizeBAD$ $g, g \cdot l$	(+) $I_{RD}\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$ $I_{SPE} \cdot totsup$ (-) $l, g\cdot sizeBAD$,$g \cdot l$
	Supergame 2	
(+) $I_{RD}\cdot\ (\delta - \delta_{RD})$, $I_{SPE}\cdot\ (\delta - \delta_{RD})$ $(\delta - \delta_{RD})$, $totsup \cdot I_{RD}$ (-) $g \cdot P, l$	(+) $I_{RD} \cdot totsup, \delta \cdot showup$ (-) $sizeBAD$, $g \cdot P$, $sizeBAD\cdot g, l$,	(+) $I_{RD}\cdot\ (\delta - \delta_{RD})$, $I_{SPE}\cdot\ (\delta - \delta_{RD})$ $(\delta - \delta_{RD})$, $totsup \cdot I_{RD}$ (-) $g \cdot P, g\cdot\ sizeBAD$, l
	Supergame 3	
(+) $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $g\cdot\ (\delta - \delta_{RD})$, $m \cdot I_{RD}$, $(\delta - \delta_{RD})$, ... (-) $\delta \cdot p, l, P \cdot showup, ...$	(+) $matching, P \cdot totsup$, ... (-) $sizeBAD$, $I_{SPE} \cdot P$ $sizeBAD\cdot I_{RD}, l$, ...	(+) $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, $P \cdot m$, $m\cdot\ (\delta - \delta_{RD})$, $m \cdot I_{RD}$, ... (-) $\delta \cdot P, l\cdot sizeBAD$, ...
	Supergame 4	
(+) $\delta\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, $I_{SPE}\cdot\ (\delta - \delta_{RD})$, (-) l	(+) (-) $sizeBAD$, l	(+) $\delta\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $I_{RD}\ (\delta - \delta_{RD})$ (-) l
	Supergame 5	
(+) $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $\delta\cdot(\delta - \delta_{RD})$, $g\cdot\ (\delta - \delta_{RD})$, $m\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, ... (-) $\delta \cdot P, g \cdot P, l,\ P \cdot showup, ...$	(+) (-) $sizeBAD$, l, $sizeBAD\ \cdot l$	(+) $P \cdot m, I_{SPE}\cdot\ (\delta - \delta_{RD})$, $\delta\cdot(\delta - \delta_{RD})$, $l \cdot m$, $m\cdot(\delta - \delta_{RD})$ (-) $P \cdot I_{RD}, g \cdot P, m, g \cdot m, , \delta \cdot l$, ...
	Supergame 6	
(+) $m\cdot\ (\delta - \delta_{RD})$, $\delta\cdot\ (\delta - \delta_{RD})$, $g\cdot\ (\delta - \delta_{RD})$, $I_{RD} \cdot \delta$, $l \cdot showup$, ... (-) $I_{SPE} \cdot P, l, P \cdot showup, I_{SPE} \cdot l$	(+) (-) $sizeBAD$, l	(+) $m\cdot\ (\delta - \delta_{RD})$, $\delta\cdot\ (\delta - \delta_{RD})$, $\delta \cdot m, g\cdot\ (\delta - \delta_{RD})$, $I_{RD} \cdot \delta$,.. (-) $P \cdot I_{RD}, I_{SPE} \cdot P, I_{SPE} \cdot l$, $m \cdot showup, P \cdot showup, ...$
	Supergame 7	
(+) $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $\delta\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, $g\cdot\ (\delta - \delta_{RD})$, I_{RD} $I_{RD} \cdot \delta, I_{RD} \cdot matching$ (-) $g \cdot P, l$	(+) (-) $sizeBAD$	(+) $I_{SPE}\cdot\ (\delta - \delta_{RD})$, $(\delta - \delta_{RD})$, $\delta\cdot\ (\delta - \delta_{RD})$ (-) $sizeBAD$, $g \cdot P, l\cdot\ sizeBAD$

Notes. Parameters are ordered based on the absolute magnitude of the estimates. Ledend: (+) and (-) indicate the sign of the estimates.

The area under the "Receiver Operating Characteristics" curve (ROC) - which varies within the interval [0,1], where 0 is the performance of a random classifier and 1 of a perfect classifier - measures the ability of the classifier to distinguish between cooperators and defectors [14]. The misclassification rate quantifies the intensity of misclassification irrespective of the misclassification error type (false positives or false negatives), while the "Precision" and the "Recall" indicators are more focused on the ability of the classifier to identify cooperators, our target of interest: the 'Precision' indicator measures how accurate positive predictions ($\hat{Y} = 1$) are, that is the probability to correctly detect positive values, while the 'Recall' indicator measures the coverage of the actual positive sample ($Y = 1$) achieved by the classifier [15].

Looking at Tables 4 and 5 and Figures 1[16], we can see that all the three models reach almost the same level of prediction accuracy across different supergames and that the prediction accuracy of the model increases as subjects move to later supergames and gain some experience. It further emerges from Table 6 and Figures 1 an asymmetry in terms of prediction accuracy along the SGE or RD equilibrium dimensions: the ML algorithm seem to produce classifiers that are more accurate - especially in terms of models' ability to detect cooperators - in treatments where game parameters are such that cooperation is sustainable in equilibrium [17].

The same evidence emerges if we collapse our observations at the treatment level and we look at how predicted probabilities estimated at the treatment level, which we interpret as the treatment-specific 'predicted cooperation rate', compare to actual observed rates of cooperation (see Tables 7 and 8 and Figure 2). The correlation increases over supergames and is steadily higher for treatments where

[14]The ROC curve summarizes the trade-off between the True Positive rate (TPR) and the False Positive rate (FPR), where $TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, TP=positives correclty predicted as positives, FN=positives incorrectly predicted as negatives, FP=negatives incorrectly predicted as positives and TN=negatives correctly predicted as negatives.

[15]Precision = $\frac{TP}{(TP+FP)}$; Recall = $TPR = \frac{TP}{(TP+FN)}$

[16]Figures 1 are based on prediction data obtained from model (1) where $\delta - \delta_{RD}$ is included among regressors. The other prediction models yield qualitatively similar results, which are reported in the Appendix (Section A.2)

[17]Figure 1 (panel a) shows that the overall ability of the algorithm to discriminate between cooperators and defectors increases over supergames, with a steadily higher disciminatory ability over observations from treatments where cooperation is sustainable as a SPE or RD equilibrium. Figure 1 (panel b) shows a modest increase in model precision over supergames, which appears to be mostly driven by an increasing ability in accurately detecting cooperators over treatments where cooperation is an equilibrium: the model ability to predict positive occurrences over treatments where cooperation is not an equilibrium, instead, soon decays to zero. A similar dynamics is shown by Figure 1 (panel c), where we observe that the coverage of the actual cooperators' sample soon decays to zero in treatments where cooperation is not an equilibrium, while remains somewhat stable, fluctuatiing around higher values, in treatments where cooperation is an equilibrium.

cooperation is an equilbrium, compared to those where it is not.

Table 4: Classification accuracy: Round-1 Cooperation – Supergame 1

	Area under ROC Curve	Misclassification (c=0.5)	FP (c=0.5)	FN (c=0.5)	Precision (c=0.5)	Recall (c=0.5)
(1) $(\delta - \delta_{RD})$ **model**	0.64	39%	20%	19%	0.60	0.62
(2) $sizeBAD$ **model**	0.64	39%	20%	19%	0.60	0.62
(3) Unconstrained model	0.64	39%	20%	19%	0.60	0.62

Table 5: Classification accuracy: Round-1 Cooperation – Supergame 7

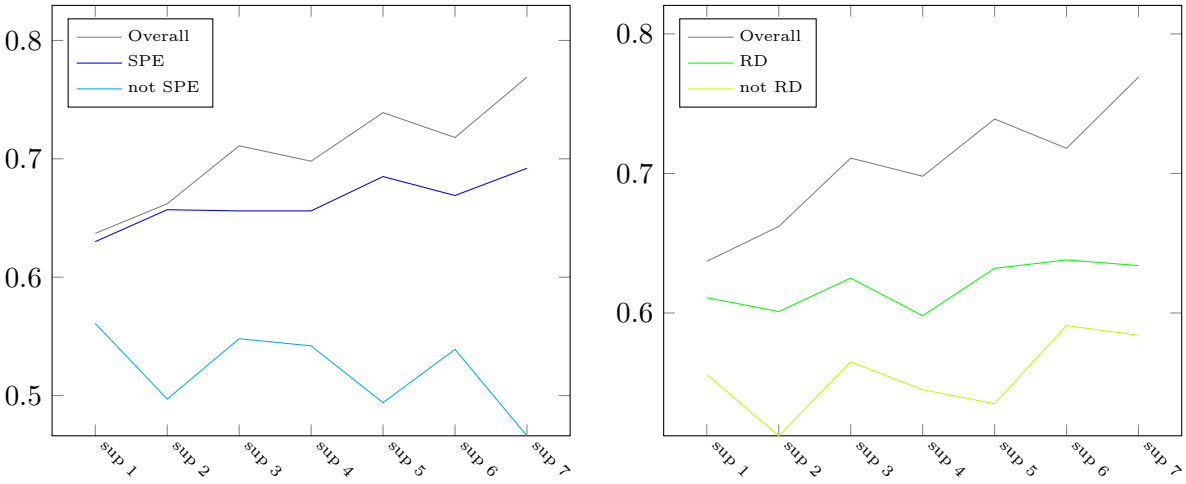
	Area under ROC Curve	Misclassification (c=0.5)	FP (c=0.5)	FN (c=0.5)	Precision (c=0.5)	Recall (c=0.5)
(1) $(\delta - \delta_{RD})$ **model**	0.77	27%	12%	15%	0.69	0.61
(2) $sizeBAD$ **model**	0.78	27%	11%	16%	0.69	0.62
(3) Unconstrained model	0.77	27%	12%	15%	0.68	0.62

Table 6: Classification accuracy: SPE vs. Not SPE treatments

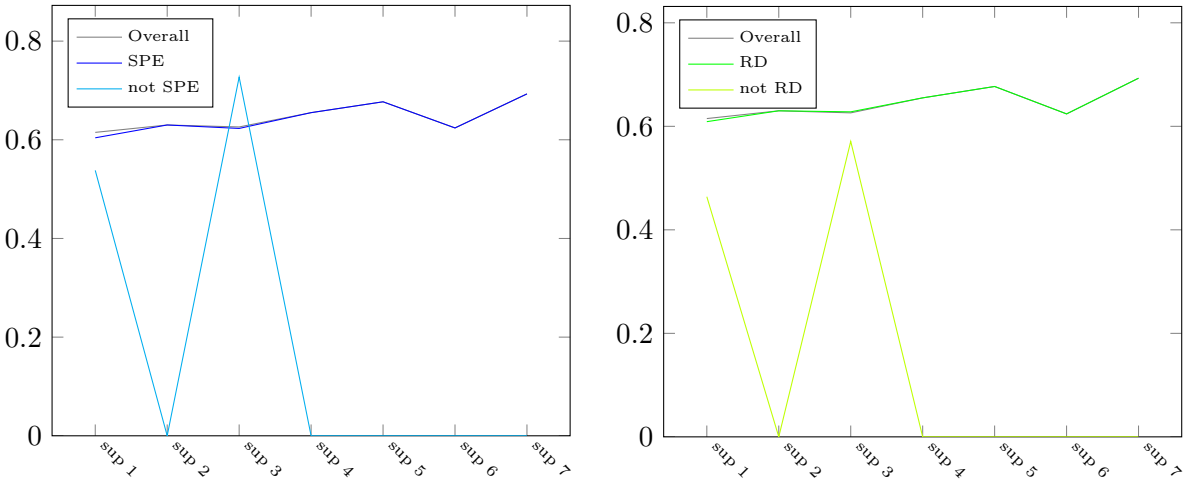
	SPE					**Non-SPE**				
	ROC	**Misscl.**	**Corr.**	**FP**	**FN**	**ROC**	**Misscl.**	**Corr.**	**FP**	**FN**
Supergame 1										
(1) $\delta - \delta_{RD}$ **model**	0.63	39%	0.56	29%	10%	0.55	40%	0.29	2%	38%
(2) $sizeBAD$ **model**	0.64	39%	0.54	29%	10%	0.56	40%	0.14	2%	38%
(3) Unconstr. model	0.63	39%	0.56	29%	10%	0.55	40%	0.23	2%	38%
Supergame 7										
(1) $\delta - \delta_{RD}$ **model**	0.69	33%	0.73	17%	16%	0.58	13%	0.05	0%	13%
(2) $sizeBAD$ **model**	0.72	33%	0.71	16%	17%	0.58	13%	.	0%	13%
(3) Unconstr. model	0.70	33%	0.74	16%	17%	0.60	13%	0.21	0%	13%

Notes. ROC = Area under the ROC curve, Misscl. = Misclassification, Corr. = Correlation between predicted and observed cooperation rate, FP = False Positives, FN = False Negatives.

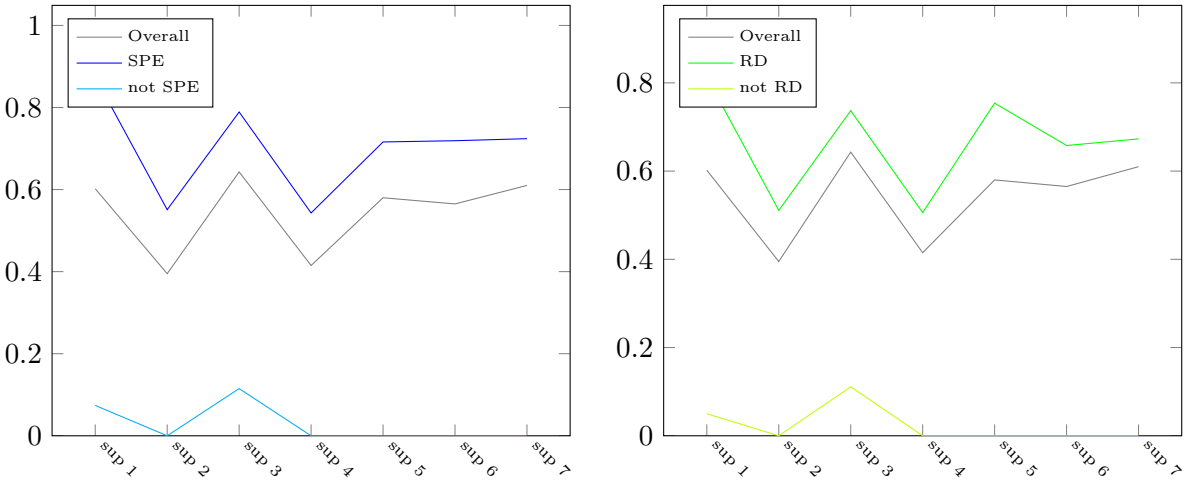
Figure 1: Classification Accuracy metrics over Supergames 1 to 7



(a) Area under the ROC curve: Supergame 1 to 7



(b) Precision: Supergame 1 to 7



(c) Recall: Supergame 1 to 7

Table 7: Prediction accuracy: Round-1 Cooperation – Supergame 1

	Correlation Predicted vs. Actual Cooperation Rate				
	Overall	SPE	Not SPE	RD	Not RD
(1) $\delta - \delta_{RD}$ **model**	59%	56%	29%	52%	28%
(2) $sizeBAD$ **model**	55%	54%	14%	50%	24%
(3) Unconstrained model	58%	56%	23%	50%	28%

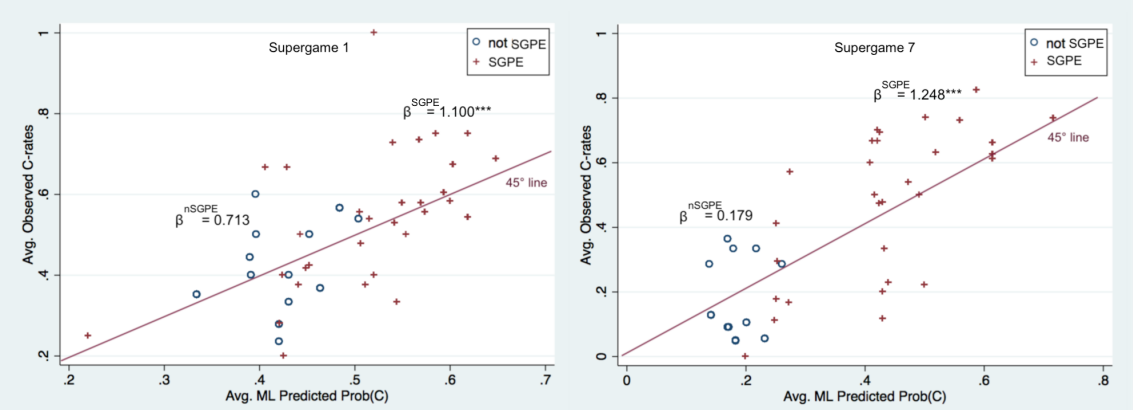
Table 8: Prediction accuracy: Round-1 Cooperation – Supergame 7

	Correlation Predicted vs. Actual Cooperation Rate				
	Overall	SPE	Not SPE	RD	Not RD
(1) $\delta - \delta_{RD}$ **model**	80%	73%	5%	60%	33%
(2) $sizeBAD$ **model**	78%	71%	. %	60%	32%
(3) Unconstrained model	81%	74%	21%	61%	44%

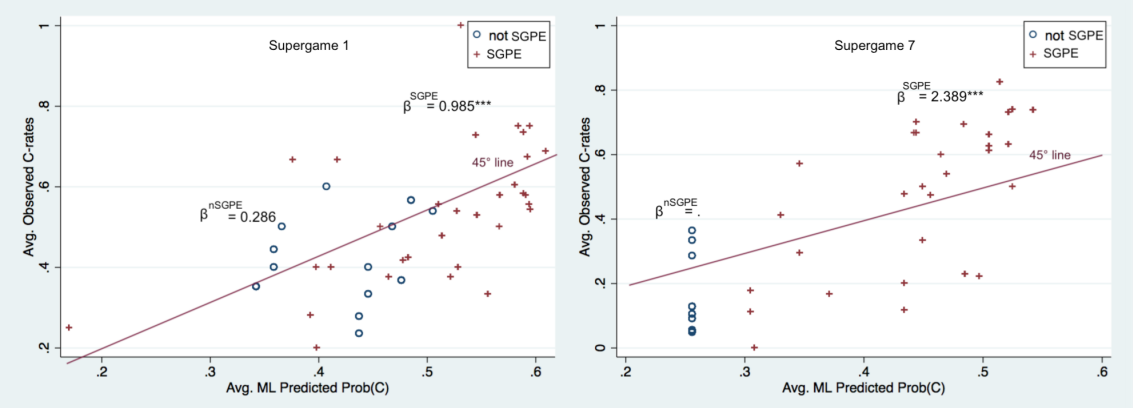
Assessing which of the two indices ($\delta - \delta_{RD}$) and $sizeBAD$ performs best in the prediction task is not straightforward. In terms of 'parsimony', that is the proximity between the model structure selected by the algorithm and the structure implied by theory - based solely on the indicator itself - models fed with $sizeBAD$ among the inputs seem to perform better [18]. However, in unconstrained model, in which we include both the indicators ($\delta - \delta_{RD}$) and $sizeBAD$ among inputs letting the algorithm free to select which one is the most useful for prediction, we observe that the ML algorithm is more likely to select the ($\delta - \delta_{RD}$) among the most relevant predictors (see Table 3). Another argument in favor of ($\delta - \delta_{RD}$) comes from the analysis on the correlation between predicted and observed cooperation rates, where models trained including ($\delta - \delta_{RD}$) slightly outperform model trained including $sizeBAD$, especially in treatments where cooperation is not sustainable as a SPE or RD equilibrium.

[18] $sizeBAD$-models are more parsimonious and achieve the same level of prediction accuracy reached by the other models, which are obtained through more complex model structures, suggesting that the $sizeBAD$ indicator alone is able to capture a great share of the variability needed to produce accurate predictions

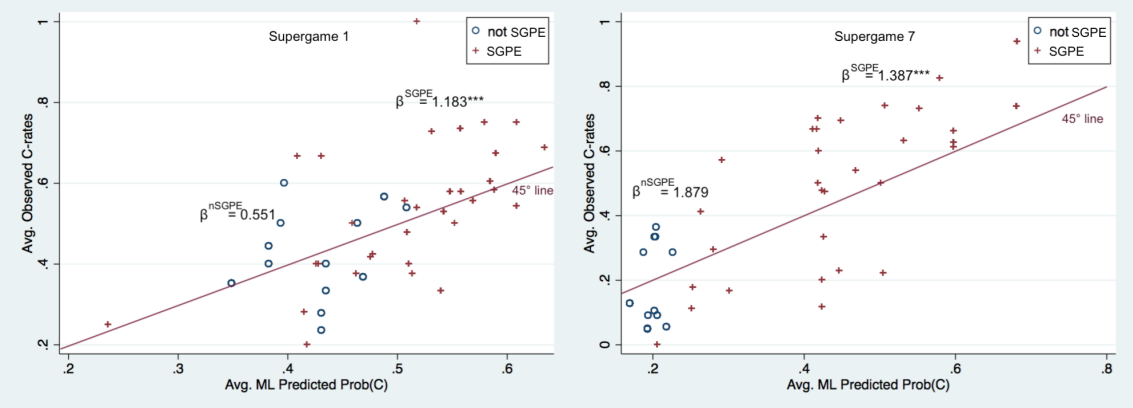
Figure 2: Average ML Predicted vs. Observed Cooperation Rates: Supergame 1 and 7



(a) $\delta - \delta_{RD}$ model



(b) $sizeBAD$ model



(c) Unconstrained model

Notes. The unit of observation is the treatment. β coefficients reported are obtained by separately estimating - for SPE and non SPE treatments - the following linear regression model $Y[Observed\ CoopRate] = \beta_0 + \beta_1 \cdot X[Predicted\ CoopRate]$. We report estimated coefficients and statistical significance of coefficients β_1.

4 Theoretical Framework

The evidence brought by the meta-analysis could be consistent with the idea that the main drivers for cooperation in contexts where cooperation is not sustainable as an equilibrium are essentially *non-strategic*. For this reason, models trained using only the information on environemntal game parameters, which capture solely the intensity of *strategic* incentives for cooperation, perform worse on the ground of the ability to detect cooperators when applied to contexts in which cooperation cannot be supported in equilibrium under standard assumptions.

This narrative would also provide a comprehensive explanation to the (relatively scarce) evidence coming from the experimental literature studying the predictive power of individual characteristics, in particular social preferences, in infinitely repeated PDs, which we will now discuss more in details.

In this framework, the behavior of individuals who exhibit a *non-strategic taste* for cooperation could be explained by the presence of some forms of social preferences, which, if strong enough, can alter incentives for cooperation favoring the emergence of cooperative equilibria even in situations where the parameters of the game are such that for a purely self-interested and payoff-maximizing individual cooperation would never be sustainable neither as a SPE nor as a RD equilibrium. These individuals would then be willing to start by playing a cooperative strategy irrespective of the environmental characteristics of the game, thus independently from the presence and the strength of strategic incentives to cooperate.

The rest of individuals, instead, would only exhibit a *strategic taste* for cooperation, according to which they would initiate a cooperative strategy only if they consider cooperation to be profitable from a strategic point of view, based on their expectations on how many individuals in the population would also cooperate, either because they rationally internalize that the environemntal parameters of the game are such that cooperating would be profitable in the long run, or because of non-strategic reasons.

In this framework, when the parameters of the game are such that cooperation is not sustainable as a long-run equilibrium, we would observe cooperation only by the first group of individuals and by the fraction of individuals from the second group who have at least a small positive expec-

tation on the fraction of individuals in the population who would be motivated to cooperate by non-strategic reasons. This prediction would be consistent with what was originally postulated by Kreps et al. [1982], who focused on the case of finitely repeated PDs and rationalized the emergence of cooperation in a context where cooperation is never sustainable as an equilibrium for rational self-interested individuals: according to Kreps et al. [1982] it would be sufficient to assume that players have incomplete information concerning preferences for cooperation in the population - so that players assign a small positive probability to the possibility that their opponent might choose to cooperate because he/she 'enjoys cooperation' - in order to produce a sequential cooperative equilibrium.

Different models of social preferences could be used to model how individuals map the monetary payoffs of a Prisoner Dilemma (PD) into utilities when deciding over their actions and strategies, and to show that, we refer to a modified version of the canonical 2x2 PD matrix, where the utilities associated to each action are displayed instead of crude payoffs, see Table 9.

If we assume players are self-interested and rational, the mapping of payoffs through the utility function does not affect the structure of the matrix, and, under the most simple and standard assumption of linearity $U_i(A_i, A_j) = \pi_i(A_i, A_j)$ [19], the utility that subjects assign to each of the possible actions will exactly correspond to their own monetary payoff associated to that action, see Table 9 (panel a). If both players are self-interested and rational, (Defect, Defect) will be the unique Nash Equilibrium (NE) of the stage game and the same will hold when the game is repeated an infinite time of times but the parameters of the game are such that cooperation would not be sustainable as a SPE or a RD equilibrium. Instead, if players exhibit some forms of social preferences, this would affect the mapping of game's monetary payoffs into utilities (as shown also by Duffy and Muñoz-García [2012]) , possibly allowing for other equilibria even in the stage game.

When accounting for social preferences, we assume players exhibit social preferences à la Charness and Rabin [2002]. In the two-players case, in absence of reciprocity concerns, the utility of the

[19] $U_i(A_i, A_j)$ is the utility subject i gets based on his own action A_i and the action of the other person in the pair A_j and $\pi_i(A_i, A_j)$ is the payoff subject i realizes based on his own action A_i and the action of the other person in the pair A_j.

individual i in the pair would be given by:

$$U_i(A_i, A_j) = f(\pi_i(A_i, A_j), \pi_j(A_i, A_j)) = (\beta_i r + \alpha_i s)\pi_j + (1 - \beta_i r - \alpha_i s)\pi_i$$

where

$r = 1$ if $\pi_i > \pi_j$, and $r = 0$ otherwise;

$s = 1$ if $\pi_i < \pi_j$, and $s = 0$ otherwise;

so that the utility of individual i can also be expressed as:

$$\begin{cases} if\ \pi_i = \pi_j \longrightarrow U_i = \pi_i \\ if\ \pi_i > \pi_j \longrightarrow U_i = \beta_i \pi_j + (1 - \beta_i)\pi_i \\ if\ \pi_i < \pi_j \longrightarrow U_i = \alpha_i \pi_j + (1 - \alpha_i)\pi_i \end{cases}$$

This framework mirrors the behavioral model adopted by Bruhin et al. [2018] to shape social preferences in absence of reciprocity concerns, which is itself inspired by the behavioral models developed by Charness and Rabin [2002] and Fehr and Schmidt [1999]. Parameters β_i and α_i measure the weights assigned by player i to the payoff of the other player both in a situation of advantageous and disadvantageous inequality, and based on the values of these two paramters it is possible to classify individuals with different types os social preferences.

When both $\beta_i = 0$ and $\alpha_i = 0$ inviduals are purely selfish and do not show any forms of social preferences, caring exclusively about their own payoff, irrespective of their relative position in the pair.

When both $\beta_i > 0$ and $\alpha_i > 0$, instead, individuals are altruistic and always care about the payoff of the other player no matter what is their relative position in the pair, showing concerns both for the maximization of the payoff of the worst-off player and for efficiency. An increase in both β_i and α_i signals an increase in the weight player i attaches to the social good, as compared to this own material payoff. When, instead, β_i increases and α_i decreases, or more in general the ratio $\frac{\beta_i}{\alpha_i}$ increases, this signals that player i puts relatively more weight to the maximization of the payoff of the worst-off player, and less to the maximization of the total surplus.

When $\beta_i > 0$ but $\alpha_i < 0$, individuals are inequality averse, which implies they are behindness averse but do care about the payoff of the other player when they are better off. Based on the relative size

of parameters | α_i | and | β_i |, individuals would either care more about advantageous inequality than disadvantageous inequality (| α_i |<| β_i |) or viceversa (| α_i |>| β_i |), where the latter represents the case originally studied by Fehr and Schmidt [1999].

When both $\beta_i < 0$ and $\alpha_i < 0$ individuals are spiteful and always attach a negative weight to the payoff of the other player, irrespective of their relative position in the pair.

If we assume players have some forms of social preferences ($\beta_i \neq 0$ and $\alpha_i \neq 0$), these preferences will have an impact on how players map payoffs into utilities when playing a PD, see Table 9 (panel b).

Table 9: Prisoners' Dilemma Row Player's Utilities - C&R

General

	C	D
C	$U_i(C,C)$	$U_i(C,D)$
D	$U_i(D,C)$	$U_i(D,D)$

Social Preferences à la Charness and Rabin (2002)

	C	D
C	$U_i(C,C) = R$	$U_i(C,D) = \alpha_i T + (1-\alpha_i)S$
D	$U_i(D,C) = \beta_i S + (1-\beta_i)T$	$U_i(C,C) = P$

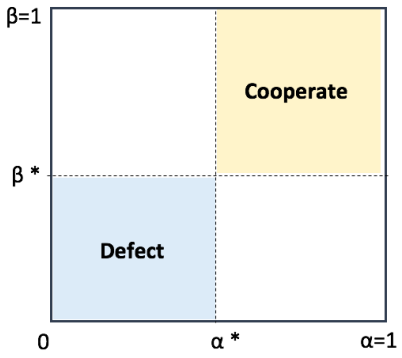
If we assume players have perfect information about social preferences, we have that under some circumstances a cooperative equilibrium can arise even in the One Shot stage game interaction. Indeed, under some circumstances, Cooperation will be a dominant strategy for both players, which will result in an efficient (Cooperate, Cooperate) Equilibrium.

$$\begin{cases} Cooperate\ is\ BR\ to\ Cooperate \longrightarrow when\ \beta_i > \beta_i^* = \frac{(T-R)}{(T-S)} \\ Cooperate\ is\ BR\ to\ Defection \longrightarrow when\ \alpha_i > \alpha_i^* = \frac{(P-S)}{(T-S)} \end{cases}$$

In order to have Cooperation as a best response to Cooperation, intuitively, we need player i to have a high enough β_i, which implies player i cares enough about the other player's payoff when $\pi_i > \pi_j$, so to compensate the loss in terms of higher material payoff he could have obtained by means of a unilateral defection. When $(T-R) < (T-S)$, which is always the case in a canonical PD where $T > R > P > S$, the threshold value β_i^* will be bounded between 0 and 1, which implies the

condition will be met only when players have an high enough degree of concern for others ($\beta_i > \beta_i^*$). Similarly, in order to have Cooperation as a best response to Defection, we need player i to have a high enough α_i, which implies player i cares enough about the other player's payoff even when $\pi_i < \pi_j$, so to compensate the loss in terms of material payoff he incurs as a consequence of his opponent's unilateral defection. When $(P-S) < (T-S)$, which is always the case in a canonical PD where $T > R > P > S$, the threshold value α_i^* will be bounded between 0 and 1, which implies the condition will be met only when players have an high enough degree of concern for others ($\alpha_i > \alpha_i^*$).

Figure 3: Players Pure Dominant strategies



Therefore, if both players are selfish and self-interested ($\beta_i = \alpha_i = 0$) or exhibit low concerns the social welfare ($\beta_i < \beta_i*$ and $\alpha_i < \alpha_i^*$), the unique NE of the stage game will be (Defect, Defect). If, instead, both players have strong enough concerns for the social welfare, the unique NE of the stage game will be (Cooperate, Cooperate). It is therefore possible to observe cooperative outcomes even in absence of any scope for future interactions.

When we move to the context of infinitely repeated PDs, players are called to play the same stage game for an indefinite number of times with the same partner and every player i discounts the flow of his future payoffs according to a discount factor $0 < \delta_i < 1$. In this context, even in absence of social preferences, the outcome (Cooperate, Cooperate) can be sustained as an equilibrium if players are sufficiently patient, as predicted by the Folk theorem (Fudenberg and Maskin [1986]).

This prediction holds whenever:

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

$$\delta_i > \delta^{SPE} = \frac{(T-R)}{(T-P)}$$

where δ^{SPE} is the threshold value that makes a player indifferent between playing a grim strategy - where the player starts by cooperating in the first round and then keeps cooperating until a defection is observed, switching to defection ever after - and an always-defect strategy, under the assumption that the opponent plays grim.

Under the assumption of both players having strong enough social preferences and perfect information, mutual cooperation can be feasible as a Subgame Perfect NE (SPE) of the infinitely repeated game under a wider set of parameters. In particular, (Cooperate, Cooperate) will be sustainable as an equilibrium outcome whenever:

$$\delta_i^{SPE} : \sum_{t=0}^{\infty} \delta^t R > \beta_i S + (1-\beta_i) T + \sum_{t=1}^{\infty} \delta^t P$$

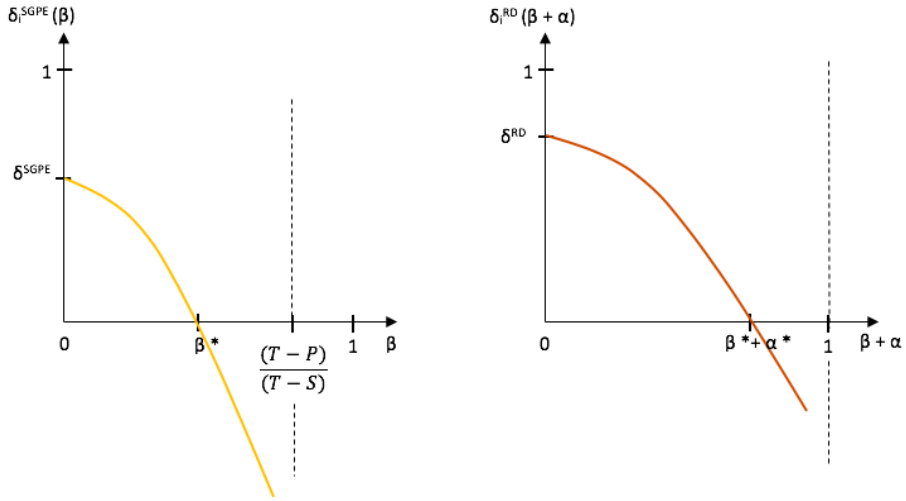
$$\delta_i > \delta_i^{SPE} = \frac{(T-R) - \beta_i (T-S)}{(T-P) - \beta_i (T-S)}$$

where δ_i^{SPE} is the threshold value that makes a player i indifferent between playing a grim strategy and an always-defect strategy.

When social preferences are absent and $\beta_i = 0$, δ_i^{SPE} and δ^{SPE} coincide. For positive values of β_i, below the threshold β_i^*, $\delta_i^{SPE} < \delta^{SPE}$, given that δ_i^{SPE} is decreasing in β_i. For high values of β_i, above the threshold β_i^*, $\delta_i^{SPE} < 0$, which implies the condition $\delta_i > \delta_i^{SPE}$ is always met and player i would always prefer to play a grim strategy.

The threshold value δ^{SPE} is obtained assuming the player i is choosing what strategy would be best to play, assuming the opponent is playing the cooperative grim strategy. Accounting for the strategic uncertainty arising from not knowing what strategy the opponent will be playing, we can still observe the outcome (Cooperate, Cooperate) being sustainable as an equilibrium among

Figure 4: δ_i^{SPE} and δ_i^{RD} as a function of β_i and $\beta_i + \alpha_i$



self-interested players if:

$$\delta_i > \delta^{RD} = \frac{T - S - R + P}{T - S}$$

where δ^{RD} is the value that makes playing grim a risk dominant strategy to the other player randomizing 50-50 between the two grim and always defect strategies.

$$\delta_i^{RD} : \frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t R\right] + \frac{1}{2}\left[S + \sum_{t=1}^{\infty} \delta^t P\right] > \frac{1}{2}\left[T + \sum_{t=1}^{\infty} \delta^t P\right] + \frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t P\right]$$

Under the assumption of both players having strong enough social preferences and perfect information, mutual cooperation can be feasible as a Risk Dominant (RD) equilibrium of the infinitely repeated game under a wider set of parameters. In particular, (Cooperate, Cooperate) will be sustainable as an equilibrium outcome whenever:

$$\delta_i > \delta_i^{RD} = \frac{T - S - R + P - (\beta_i + \alpha_i)(T - S)}{T - S - (\beta_i + \alpha_i)(T - S)}$$

When social preferences are absent, so that $\beta_i = \alpha_i = 0$, $\delta_i^{RD} = \delta^{RD}$. For positive values of β_i, α_i such that $\beta_i + \alpha_i < \beta_i^* + \alpha_i^*$, $\delta_i^{RD} < \delta^{RD}$ since δ_i^{RD} is decreasing in $\beta_i + \alpha_i$. For positive values of β_i, α_i such that $\beta_i + \alpha_i > \beta_i^* + \alpha_i^*$, $\delta_i^{RD} < 0$, which implies the condition $\delta_i > \delta_i^{RD}$ always holds and player i would always find profitable to play a grim strategy.

These conclusions are not specific to the choice of modeling social preferences using a model à la Charness and Rabin [2002]. Indeed, we could also model social preferences relying on the original model by Fehr and Schmidt [1999] and we would obtain qualitatively equivalent results [20]. In this framework, if we compare the behavior of indivduals with strong enough altruistic preferences (with positive and large α_i and β_i) - the *Strongly Altruistic* types - with the behavior of others - the *Not Strongly Altruistic* types - we expect it to differ when cooperation is not an equilibrium, and to be more comparable when cooperation is sustainable in equilibrium. In particular, we expect:

Hypothesis 1. Strongly Altruistic types to cooperate, on average, more than the Not Strongly Altrusitic types when cooperation is not sustainable in equilibrium;

Hypothesis 2. Individuals' social preference type to be a good predictor of cooperation in treatments where cooperation is not an equilibrium, but not necessarily whencooperation is an equilibrium.

Hypothesis 3. Groups with a higher concentration of Strongly Altruistic types to reach higher levels of cooperation when cooperation is not an equilibrium, through a beliefs updating mechanism; instead, we expect no striking difference across groups with a different concentration of Strongly Altruistic types when cooperation is sustainable as an equilibrium.

Hypothesis 4. Strongly Altruistic types to be less sensitive to changes in the environmental incentives to cooperation,depending on game paramenters.

Except for a few works, the evidence on motives behind cooperation in infinitely repeated PDs across strategically different scenarios is not abundant. Reuben and Suetens [2012] study an indefinitely repeated PD where cooperation is not sustainable in equilibrium employing the strategy-method to

[20]For a discussion, see the Appendix (Section A.3).

disentangle strategically and non-strategically motivated behavior and to identify to what extent strategically motivated individuals are responsible for observed cooperative patterns. By adopting a sequential design where both players can submit their actions conditional on whether or not the round they are playing is the last one, Reuben and Suetens [2012] are able to study the end-game effect and to distinguish strategically from non-strategically motivated second movers: they find that cooperation is greater when the round played is not the last one, which suggests a prevalent end-game effect and a large scope for strategically motivated cooperation, although a role for non-strategically motivated cooperation driven by individual preferences also emerges. They further document that individuals' motivation to cooperate over time is stable, which suggests individuals choose to cooperate either for strategic or non-strategic considerations and behave consistently over time.

5 Experimental Design

We test whether our hypotheses are supported by empirical evidence through a novel experimental design, through which we are able to collect information on both:

- subjects' social preferences, as to distinguish the *Strongly Altruistic* types from the rest;
- subjects' actual behavior in infinitely repeated PDs

The proposed experimental design is divided into two parts: in Part 1 we measure subjects' social preferences and collect some information on subjects' individual characteristics through a questionnaire; in Part 2, we observe subjects' behavior in Infinitely Repeated PDs, eliciting subjects' beliefs on the share of Round-1 cooperators across supergames.

At the beginning of the experiment, subjects learn about the two-parts structure of the experiment and that they will be informed of their earnings and paid only at the end of the last part of the experiment. Parts 1 is the same across treatments, while the design of Part 2 varies across treatments. Each subject is exposed to one treatment only.

The experiment has been programed using *Otree* and run entirely online [21]. Subjects were recruited

[21]The first wave of data collection has been run between May 16 and May 26, 2020. The second wave of data collection has been run between June 30 and July 7, 2020

from the local pool of students of the University of Bologna using ORSEE (Greiner [2004]). [22] a total of 288 subjects completed both parts of the experiment [23]. Subjects spent on average 20 minutes to take part in Part 1 of the experiment and sessions of Part 2 of the experiment lasted on average one hour and five minutes. Subjects earned on average, 15.25 Euros (payments ranged within 8.8 and 27.8 Euros), including the 4 Euros show-up fee.

Subjects never interacted with the same counterparts in Part 1 and Part 2. The structure of the experimental design is tailored to minimize the risk of spillover effects between Part 1 and Part 2 [24]. Parts 1 and 2 took place about four days apart.

Upon registration, subjects have been invited to participate in Part 1 of the experiment [25]. Only subjects who completed all tasks from Part 1 within the due date were invited to participate in Part 2, where subjects actually interacted in real time with their counterparts. Subjects were paid only at the end of Part 2, provided that they completed Part 1 within the due date and logged-in on time for Part 2, to prevent attrition and drop-out [26].

5.1 Part 1

In Part 1 we estimate subjects' social preferences paramenters relying on the procedure recently developed by Bruhin et al. [2018]. Different alternative procedures have been proposed over the past years to estimate social preferences in the literature, which mainly differ in terms of: (i) the degree flexibility in terms of the number of social preferences categories considered, which are either

[22] The experiment was pre-registered on the Open Science Framework (OSF) Registry: https://osf.io/mzt74

[23] Additional 16 participants took part in the pilot session of the experiment, run in early May 2020: data from this session are discarded from the analysis due to a slight difference in experimental procedures, namely a shorter time window between Part 1 and Part 2.

[24] Indeed, there is some evidence that having subjects playing both the infinitely repeated PDs and other games aimed to measure their social preferences could lead to contamination and spillover issues: Peysakhovich and Rand [2016], for example, by letting their subjects play a series of cooperation games, including a DG after an infinitely repeated PD with varying continuation probabilities, document that subject cooperativeness is significantly impacted by how conducive to cooperation was the environment they faced in the PD, with subjects exposed to the treatment where cooperation was an equilibrium exhibiting higher cooperativeness. This suggests that, even in absence of pure hedging or income effects that may be triggered by the multi-games structure of the experiments, contamination and across-games spillovers might still be an issue, possibly leading to attenuated or exacerbated results.

[25] Since the decision environment faced by subjects in Part 1, which will be further described in this section, is essentially non-strategic, subjects were invited to complete Part 1 whenever they wanted over a pre-defined time frame that expired some days prior to the moment when Part 2 of the experiment actually took place.

[26] Attrition from Part 1 to Part 2 of the experiment was on average equal to 6.4%. We did not register any drop-out cases during the online interactive sessions of the experiment (Part 2).

determined ex-ante or determined endogenously; and (ii) the choice to adopt a fully non-parametric or parametric approach [27]. Bruhin et al. [2018] propose a parametric approach based on a behavioral model inspired by the work by Fehr and Schmidt [1999] and Charness and Rabin [2002], which is extended to account also for positive and negative reciprocity concerns:

$$U_i(\pi_i, \pi_j) = (\beta_i r + \alpha_i s + \gamma q + \eta v)\pi_j + (1 - \beta_i r - \alpha_i s - \gamma q - \eta v)\pi_i$$

where :

$r = 1$ if $\pi_i > \pi_j$, and $r = 0$ otherwise;

$s = 1$ if $\pi_i < \pi_j$, and $s = 0$ otherwise;

$q = 1$ if player j behaved kindly toward i, and $q = 0$ otherwise (positive reciprocity);

$v = 1$ if player j behaved unkindly toward i, and $v = 0$ otherwise (negative reciprocity);

The resulting behavioral model provides a parsimonious characterization of subjects' social preferences through a vector of 4 parameters $\theta = \{\alpha, \beta, \gamma, \eta\}$, which are estimated from experimental choice. This approach does not impose any ex-ante constraints on the number or the characteristics of social preferences' types, which are endogenously determined through a Finite Mixture model.
In our case, we rely on a simplified version of their behavioral model, where only distributional preferences are considered and reciprocity concerns are not accounted for ($\gamma = 0$ and $\delta = 0$ so that $\theta = \{\alpha, \beta\}$):

$$U_i(\pi_i, \pi_j) = (\beta_i r + \alpha_i s)\pi_j + (1 - \beta_i - \alpha_i s)\pi_i$$

Indeed, Bruhin et al. [2018] claim that distributional preferences turn out to be considerably more important than reciprocity preferences, and in the context of our analysis, which is focused on 1^{st} Round choices in infinitely repeated PDs, reciprocity preferences are likely to play a minor role. We choose to rule out reciprocity concerns for a matter of simplicity but accounting for reciprocity concerns will surely provide a more complete and comprehensive picture, especially in the analysis of cooperative outcomes' survival over rounds and across supergames, and we aim to extend our analysis in this direction in a future work.
In the orginal work by Bruhin et al. [2018], social preferences parameters $\theta = \{\alpha, \beta, \gamma, \eta\}$ are estimated relying on a set of 117 binary decisions data. In each binary decision situation, subjects have

[27] For a discussion, see Bruhin et al. [2018].

to choose between two possible payoff allocations between themselves and an anonymous player j. These binary decision situations are represented by: (i) a series of 39 dictator games for identifying the parameters α and β, and (ii) a series of 78 reciprocity games for identifying γ and η. Since we are interested in estimating only distributional preferences parameters $\theta = \{\alpha, \beta\}$, we rely only on a set of 39 dictator games to obtain the binary decision choice data necessary for the estimation. This procedural difference does not deeply affect neither the results of the estimation of parameters α and β and the characterization of preferences types, nor the categorization of subjects into preference types and the quality of out-of-sample predictions [28].

In Part 1, subjects play the same set of 39 Dictator Games designed by Bruhin et al. [2018], which are shown in Table **??** in the Appendix [29]. In each dictator game, subjects play in the role of the dictator (player i), who can either increase or decrease the payoff of player j by choosing one of two possible payoff allocations, $X = (\Pi_X^i; \Pi_X^j)$ or $Y = (\Pi_Y^i; \Pi_Y^j)$. In order to identify subject's distributional preferences, governed by α and β, the cost of changing the other player's payoff systematically varies across the dictator games. The dictator games are presented in blocks and appear in random order across subjects. Subjects are paid only for one binary decision situation in which they played as player i, that is randomly selected for payment and paid at the end of the experiment [30]. The distribution of payoffs in this decision situation is paid out both to subjects playing as Person i and to the randomly matched partners, selected to play the role of the receiver (Person j).

After playing the Dictator games, subjects are asked to answer a Questionnaire, soliciting personal and demographic data (gender, age, major), numeracy ability, based on the 8-items measure developed by Weller et al. [2013], and personality traits, assessed through the big five personality

[28] see the Appendix (Section A.4) for a discussion based on the data from Bruhin et al. [2018]

[29] see (Section A.4). Please note that amounts are expressed in terms of Experimental Currency Units (ECUs). In Part 1, ECUs are converted Euros at a conversion rate such that 100 ECUs = 0.4 Euros.

[30] The choice of paying subjects only for one randomly selected choice, out of the many made throughout the experiment ("pay one" approach) can be seen as an alternative to the more traditional choice of paying subjects for all choices made ("pay all" approach). The "pay one" approach, which helps in avoiding wealth effects and issues related to hedging and bankruptcy in experiments involving multiple decisions, is increasingly gaining momentum in the experimental literature, with recent theoretical (Azrieli et al. [2018]) and empirical (Charness et al. [2016]) contributions suggesting the "pay one" approach can prove to be as effective as the "pay all" approach, or eventually better in some cases.

dimensions measured using the 44-items Big Five Inventory developed by John and Srivastava [1999].

5.2 Part 2

In Part 2 each subject plays two blocks of infinitely repeated Prisoner Dilemmas: the matrix of the monetary payoffs, in terms of Experimental Currency Units (ECUs), is shown in Figure 5.

The set of continuation probabilities faced by subjects changes across treatments. Irrespective of what treatment they are exposed to, all subjects play two blocks of 10 supergames each. In one of these two blocks, they play a series of 10 One Shot games, where the continuation probability is set equal to zero ($\delta^{'} = 0$) and is the same across treatments; in the other block, they play a series of 10 supergames with a continuation probability ($\delta^{''}$) that varies across treatments. [31]

Therefore, $\delta^{''}$ is the principal treatment variable:

- in **T1:** $\delta^{''} = 0.35$, so that cooperation is not sustainable among self-interested players neither as a SPE or a RD equilibrium;
- in **T2:** $\delta^{''} = 0.55$, so that cooperation is sustainable among self-interested players as a SPE but not as a RD equilibrium;
- in **T3:** $\delta^{''} = 0.75$, so that cooperation is sustainable among self-interested players both as a SPE and a RD equilibrium.

We adopt a perfect-stranger matching procedure across blocks and a random-matching procedure within blocks. To reduce contagion among subjects and increase the speed of convergence towards an equilibrium, we used matching groups of 4 subjects [32].

Subjects are informed of their opponent's choices and their round-payoff at the end of each round, and of their overall supergame-payoff in ECUs at the end of each supergame [33]. Subjects are paid only for the payoff they realize in one supergame per block, which will be randomly selected at the end of the experiment.

We further elicit subjects' perceptions on the share of Round-1 cooperators in their group in both blocks. In the $1^{st}, 5^{5th}$ and 10^{th} supergame of each block, subjects are asked to guess, before actually

[31] The order of $\delta^{'}$ and $\delta^{''}$ blocks is randomized across sessions to control for order effects.
[32] see the Appendix (Section A.5 for more details on the Recruitment & Matching Procedures.
[33] In Part 2, ECUs are converted Euros at a conversion rate such that 100 ECUs = 2 Euros.

playing, what is the number of subjects in their group who would start by cooperating. If one of these supergame is selected as relevant for payment and subjects' guess matches the actual fraction of cooperators in the first round of the selected supergame, subjects receive an additional fixed payment of 2 Euros.

Figure 5: PD Monetary Payoffs matrix

	C	D
C	73, 73	10, 100
D	100, 10	43, 43

$g = l = 0.9$

Table 10: Experimental Design - Differences across Treatments

	Treatment 1	Treatment 2	Treatment 3
PART 1	Pref. Elicitation	Pref. Elicitation	Pref. Elicitation
	+ Questionnaire	+ Questionnaire	+ Questionnaire
PART 2	Inf. Rep. PDs	Inf. Rep. PDs	Inf. Rep. PDs
	$\delta' = 0$; $\delta'' = 0.35$	$\delta' = 0$; $\delta'' = 0.55$	$\delta' = 0$; $\delta'' = 0.75$
	+ Belief Elicitation	+ Belief Elicitation	+ Belief Elicitation

6 Results

The analysis of the experimental data collected through Part 1 and Part 2 of the experiment allows us to test whether the hypothesis that individuals with and without a strong *non-strategic* taste for cooperation behave differently in terms of cooperative attitudes in infinitely repeated PDs, is supported by the empirical evidence.

Relying on the data collected in Part 1 of the experiment (DGs choices), we are able to retrieve:

- individual-level estimates of social preferences parameters (α_i and β_i)
- type-specific estimates of social preferences parameters (α_t and β_t), which allow us to categorize subjects into one of the three social preference types: Behindness-Averse (BA), Moderately Altruistic (MA), and Strongly Altruistic (SA) types.

Table 11 shows some summary statistics on the estimates of social preference parameters obtained at the type and at the individual level.

Provided that, as discussed in Section 2, we are interested in testing qualitative predictions based on the distance between the estimated social preference parameters and the threshold values α^* and β^* [34], we pool together Behindness-Averse (BA) and Moderately Altruistic (MA) types and rely on a binary classification that distinguishes Strongly Altruistic (SA) types from the rest.

Throughout the analysis we focus on Round1-Cooperation choices only [35].

Table 11: Summary statistics of social preference parameters estimates

Type-specific estimates			
	Behindness-Averse	Moderately Altruistic	Strongly Altruistic
α	-0.254	0.029	0.18
β	0.015	0.046	0.418

Summary of individual-specific estimates		
	Strongly Altruistic	Not Strongly Altruistic
α	0.164 [0.188]	-0.157 [0.510]
β	0.414 [0.205]	0.007 [0.303]

Notes: Type-specific estimates are obtained from the Finite Mixture Model with k=3. Individual-specific estimates are obtained through a procedure that estimates the parameters separately for each subject. For details on the estimation procedure see Bruhin et al. [2018]. Standard deviation in squared brackets.

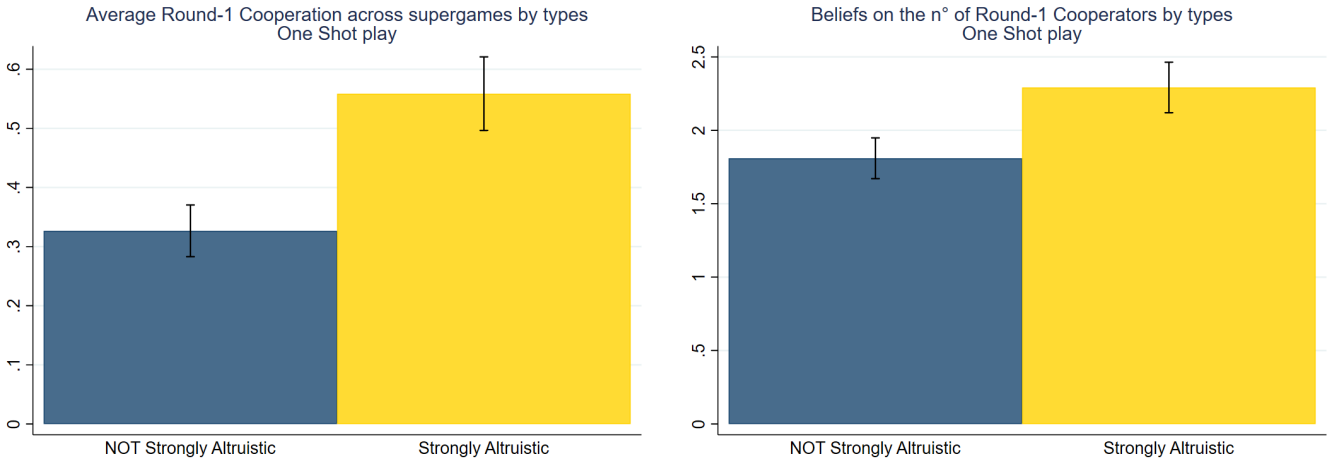
Result 1. *Strongly Altruistic types do cooperate more than others when cooperation is not sustainable as an equilibrium.*

[34] $\alpha^* = \frac{(P-S)}{(T-S)} = \frac{(43-10)}{(100-10)} \simeq 0.37$ is the threshold value that makes Cooperation become a best response to Defection in a One Shot PD game. $\beta^* = \frac{(T-R)}{(T-S)} = \frac{(100-73)}{(100-10)} \simeq 0.30$ is the threshold value that makes Cooperation become a best response to Cooperation in a One Shot PD game, and Grim a preferred strategy to Always Defect in an infinitely repeated PD irrespective of the actual value of δ.

[35] We exclude 23 subjects from the original 288 subjects sample because they behaved very inconsistently in the series of Dictator Games, therefore we were not able to identify their social preference parameters.

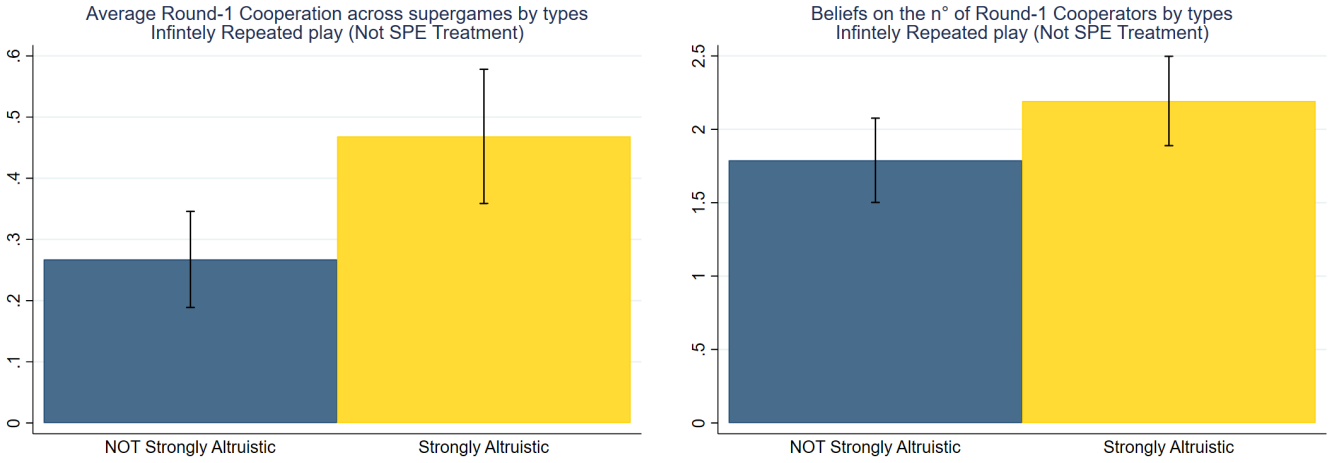
Result 1 is supported by the evidence brought by Table 12 and Figures 6 and 7, which show average cooperativeness (Left panel) and average beliefs on the number of Round-1 cooperators in the group (Right panel) across supergames.

Figure 6: Round-1 Cooperation and Beliefs across supergames - One Shot play



Notes. Data on One Shot play choices, pooling osbervations from all sessions with $\delta = 0$ (N=265).

Figure 7: Round-1 Cooperation and Beliefs across supergames - Infinitely Repeated play (Not SPE)



Notes. Data from Infinitely Repeated play choices from Treatment 1 with $\delta = 0.35$ (N=87);

Table 12: Analysis of Round-1 Cooperation and Beliefs in One Shot and Inf. Rep. play (Not SPE)

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Round-1 Cooperation				Beliefs	
Strongly Altruistic (SA)	0.224***	0.128***	0.208***	0.159**	0.494***	0.381**
	(0.0436)	(0.0369)	(0.0583)	(0.0623)	(0.114)	(0.181)
Belief		0.195***		0.130***		
		(0.0174)		(0.0389)		
Constant	0.0506	-0.296	1.366***	0.720	1.779**	4.959***
	(0.343)	(0.292)	(0.473)	(0.558)	(0.808)	(1.092)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.147	0.437	0.198	0.334	0.0838	0.178
N	265	265	87	87	265	87

Notes: Models (1)-(6): Estimated coefficients of OLS models. Models (1),(2),(5) are estimated on One Shot play choices (N=265), while Models (3),(4),(6) are estimated on Infinitely Repeated play choices from Treatment 1 (N=87). Models (1)-(4): Dependent variable = Average Round-1 Cooperation across supergames. Models (5)-(6): Dependent variable = Average Belief on Round-1 Cooperators in the group across supergames.

"Strongly Altruistic" (SA) is a dummy valued 1 if the subject is classified as a Strongly Altruistic type. "Belief" is the average belief on the number of Round-1 cooperators in the group. Indivdual controls include: age, gender, Econ major, the Big-Five personalities dimensions and a measure of numeracy.

Clustering at matching group level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Table 12 and Figures 6, 7 show that Strongly Altruistic individuals do exhibit a higher cooperative tendency than others, even after controlling for individual beliefs on the share of cooperators in the group, which have a strong and positive per-se effect on cooperation. Indeed, beliefs explain a large fraction of the variation in observed cooperation levels and - despite large within-group heterogeneity - Strongly Altruistic types show on average significantly higher beliefs, which seems to be in line with the "false consensus" effect documented by Butler et al. [2015] in trust game experiments [36].

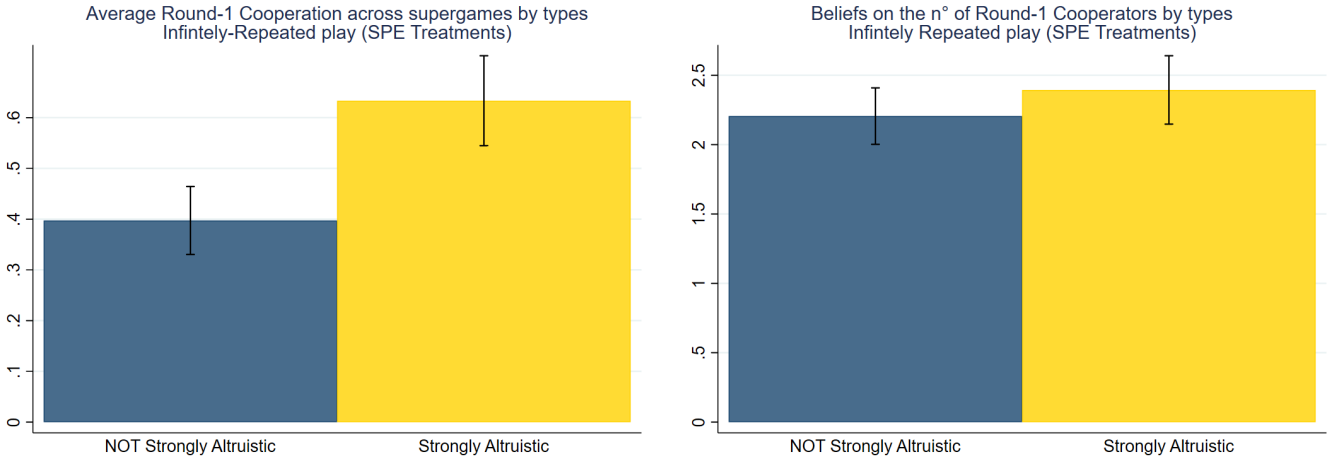
[36]Butler et al. [2015], in a trust game experiment, show that subjects tend to form their trust beliefs based on their

Result 2. *Strongly Altruistic types cooperate more than others also when cooperation is sustainable as an equilibrium. Accordingly, social preferences predict cooperation at the individual level both when cooperation is and is not sustainable as an equilibrium.*

Result 2 finds support in Figure 8 and in the results reported in Tables 13 and 14.

Figure 8 and Table 13 show that Strongly Altruistic individuals actually exhibit a higher cooperative tendency than others also in contexts where enviornmental paramenters of the game are such that cooperation is sustainable as an equilibrium. Nevertheless, in these contexts, Strongly Altruistic individuals do not substantially differ from others in terms of average beliefs on the share of cooperators in the group.

Figure 8: Round-1 Cooperation and Beliefs across supergames - Infinitely Repeated play (SPE)



Notes. Data from Infinitely Repeated play choices from Treatments 2 and 3, with $\delta = 0.55$ and $\delta = 0.75$, pooled (N=178);

Table 14, pooling observations on Infinitely Repeated play from all treatments, allows us to test whether accounting for social preference types contributes to capture a higher share of variation in cooperation levels in treatments where cooperation is not sustainable as an equilibrium. In contrast with our initial conjecture, but in line with the evidence reported earlier on the relevance of social preference types in contexts where cooperation is an equilibrium, the social preference type cate-

own trustworthiness, and that this pattern is persistent over time despite learning.

gorization has a remarkable impact on explained variation regardless of the strategic environment. Moreover, most of the effect seems to be transmitted through the beliefs channel, provided that, especially in contexts where cooperation is not an equilibrium, Strongly Altruistic types report significantly highers beliefs than others.

Table 13: Analysis of Round-1 Cooperation and Beliefs in Infinitely Repeated play (SPE)

	(1)	(2)	(3)	(4)
	Average Round-1 Cooperation		Belief	
Strongly Altruistic (SA)	0.214***	0.172***	0.204	0.204
	(0.0622)	(0.0486)	(0.191)	(0.191)
Belief		0.206***		
		(0.0178)		
Constant	0.708	0.0292	3.292**	3.292**
	(0.511)	(0.420)	(1.392)	(1.392)
Controls	Yes	Yes	Yes	Yes
R-squared	0.121	0.446	0.0351	0.0351
N	178	178	178	178

Notes: Models (1)-(4): Estimated coefficients of OLS models. All Models are estimated on Infinitely Repeated play choices from Treatments 2 and 3 pooled (N=178). Models (1)-(2): Dependent variable = Average Round-1 Cooperation across supergames. Models (3)-(4): Dependent variable = Average Belief on Round-1 Cooperators in the group across supergames.

"Strongly Altruistic" (SA) is a dummy valued 1 if the subject is classified as a Strongly Altruistic type. "Belief" is the average belief on the number of Round-1 cooperators in the group. Indivdual controls include: age, gender, Econ major, the Big-Five personalities dimensions and a measure of numeracy.

Clustering at matching group level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Table 14: Analysis of Round-1 Cooperation and Beliefs in Infinitely Repeated play

	(1)	(2)	(3)	(4)	(5)
	Average Round-1 Cooperation				Belief
SPE	0.139**	0.0567			0.442**
	(0.0630)	(0.0512)			(0.208)
Strongly Altruistic (SA)	0.206***	0.124**			0.439**
	(0.0575)	(0.0525)			(0.200)
SPE x Strongly Altruistic	0.0115	0.0572			-0.245
	(0.0860)	(0.0720)			(0.276)
Belief		0.187***	0.186***	0.184***	
		(0.0168)	(0.0172)	(0.0170)	
$\delta - \delta_{SPE}^{type}$			0.177***		
			(0.0371)		
$\delta - \delta_{RD}^{type}$				0.311***	
				(0.0643)	
Constant	0.731*	0.127	0.178	0.231	3.241***
	(0.402)	(0.346)	(0.331)	(0.334)	(1.064)
Controls	Yes	Yes	Yes	Yes	Yes
R-squared	0.139	0.409	0.411	0.412	0.0680
N	265	265	265	265	265

Notes: Models (1)-(5): Estimated coefficients of OLS models. Models (1)-(4): Dependent variable = Average Round-1 Cooperation across supergames. Model (5): Dependent variable = Average Belief on Round-1 Cooperators in the group across supergames.

"SPE" is a dummy valued 1 if the subject is exposed to a treatment where cooperation is sustainable as an equilibrium. "Strongly Altruistic" (SA) is a dummy valued 1 if the subject is classified as a Strongly Altruistic type. "Belief" is the average belief on the number of Round-1 cooperators in the group. The two indices ($\delta - \delta_{SPE}^{type}$) and ($\delta - \delta_{RD}^{type}$) measure the distance between the actual δ and the perceived - type-specific - threshold values for δ that would make cooperation sustainable as a SPE or RD equilibrium for a subject belonging to that type. Indivdual controls include: age, gender, Econ major, the Big-Five personalities dimensions and a measure of numeracy.

Clustering at matching group level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

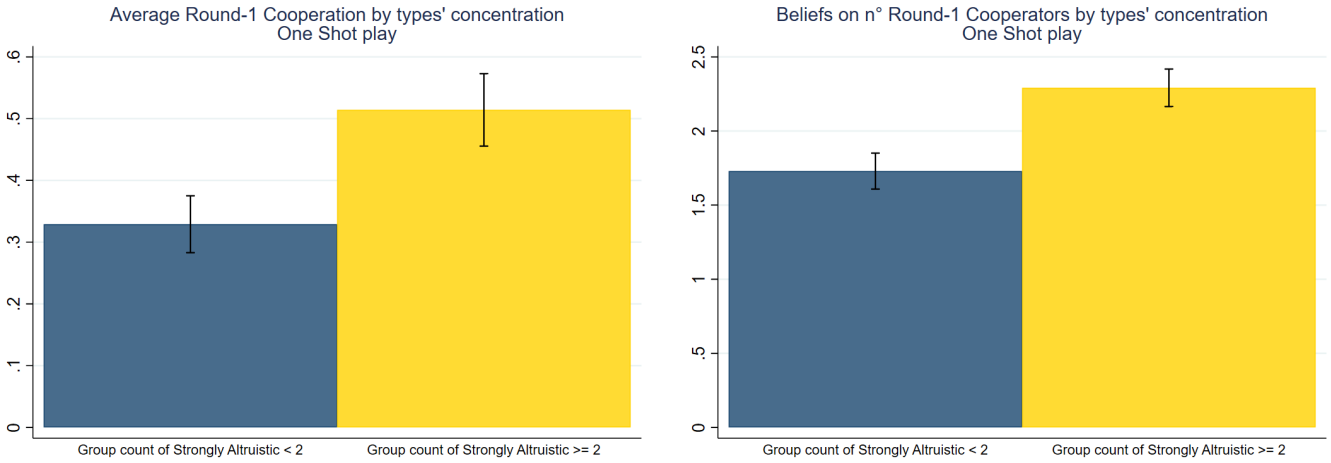
Result 3. *Groups' behavior differs based on count of Strongly Altruistic types within group in treatments where cooperation is not sustainable as an equilibrium, but not in treatments where cooperation is an equilibrium.*

Result 3 is supported by the evidence reported by Table 15 and Figures 9, 10, 11.

Figures 9, 10, 11 show average cooperativeness across supergames, conditional on the number of Strongly Altruistic types within the matching group (Left panels).

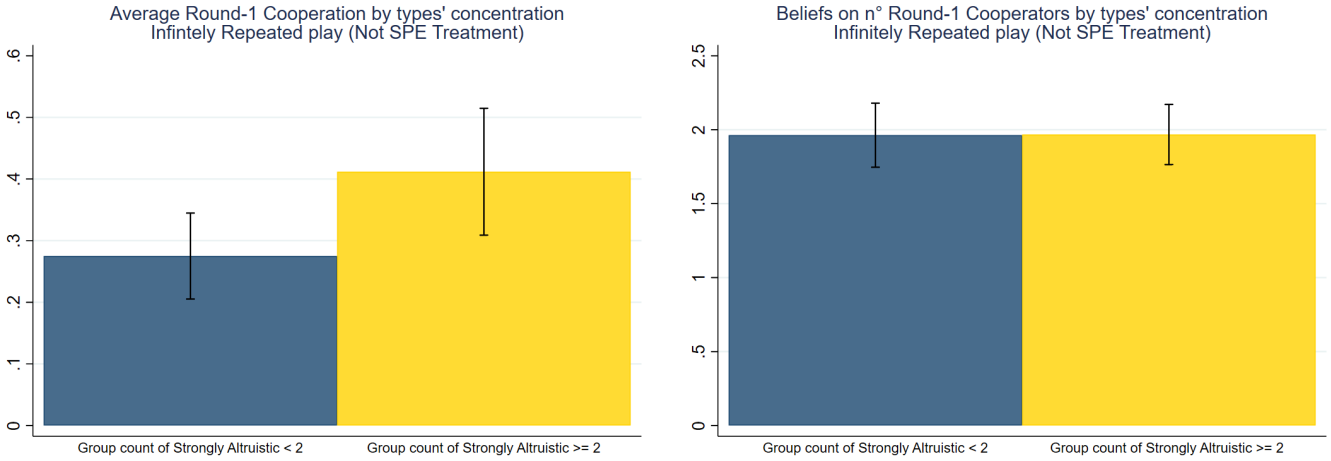
As further confirmed also by the results in Table 15, groups with a higher count of Strongly Altruistic types tend to reach a higher level of cooperation in treatments where cooperation is not sustainable as an equilibrium, including the case of One Shot interactions, but not in treatments where cooperation is an equilibrium. This evidence could be consistent with the idea that in contexts where cooperation is not an equilibrium, the presence of Strongly Altruistic types fosters cooperation even among individuals who wouldn't have cooperated otherwise, possibly through the beliefs channel, provided that beliefs are significantly higher when the concentration of Strongly Altruistic is higher. In contexts where cooperation is, instead, an equilibrium also for purely self-interested individuals, beliefs are on average higher for all and the count of Strongly Altrustic types within the group does not dramatically impact neither overall cooperativeness nor beliefs.

Figure 9: Round-1 Cooperation and Beliefs across supergames - One Shot play



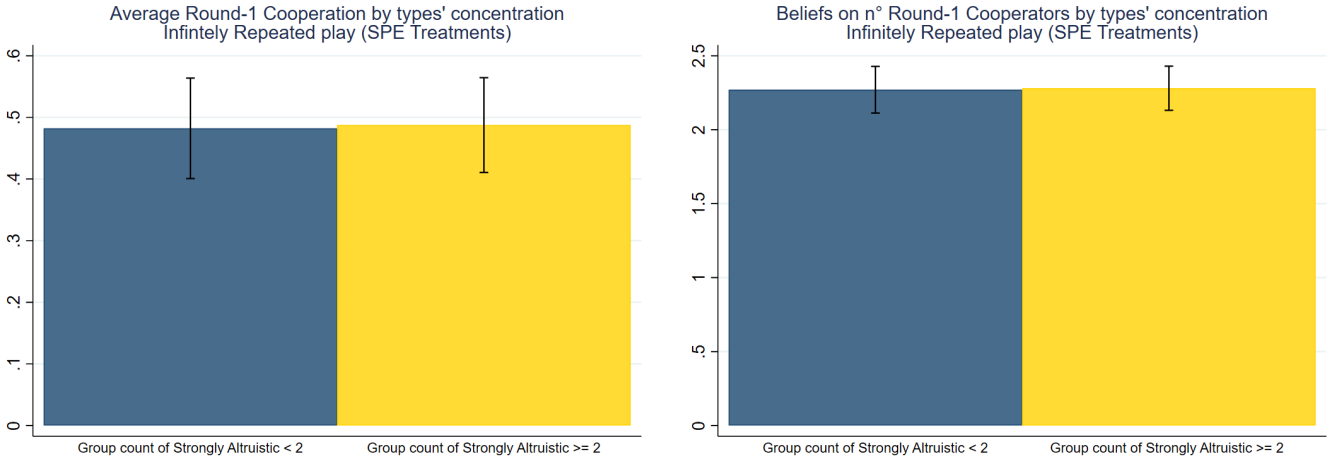
Notes. Data from One Shot play (N=265);

Figure 10: Round-1 Cooperation and Beliefs across supergames - Infinitely Repeated play (Not SPE)



Notes. Data from Infinitely Repeated play, Treatment 1 with $\delta = 0.35$ (N=87);

Figure 11: Round-1 Cooperation and Beliefs across supergames - Infinitely Repeated play (SPE)



Notes. Data from Infinitely Repeated play, Treatments 2 and 3 with $\delta = 0.55$ and $\delta = 0.75$ pooled (N=178);

Table 15: Analysis of Round-1 Cooperation and Beliefs: One Shot and Inf. Rep. play

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Round-1 Cooperation			Beliefs		
Count of SA ≥ 2	0.0615*	0.167***	-0.0109	0.457***	-0.161	-0.0541
	(0.0360)	(0.0592)	(0.0518)	(0.152)	(0.218)	(0.228)
Belief	0.202***	0.146***	0.213***			
	(0.0188)	(0.0375)	(0.0180)			
Strongly Altruistic (SA)				0.287***	0.442**	0.228
				(0.107)	(0.192)	(0.205)
Constant	-0.233	0.621	0.232	1.755**	4.956***	3.325**
	(0.302)	(0.542)	(0.430)	(0.815)	(1.111)	(1.399)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.410	0.335	0.401	0.133	0.183	0.0356
N	265	87	178	265	87	178

Notes: Models (1)-(6): Estimated coefficients of OLS models. Models (1)-(3): Dependent variable = Average Round-1 Cooperation across supergames. Models (4)-(6): Dependent variable = Average Belief on Round-1 Cooperators in the group across supergames.

"Strongly Altruistic" (SA) is a dummy valued 1 if the subject is classified as a Strongly Altruistic type. "Belief" is the average belief on the number of Round-1 cooperators in the group. "Count of SA ≥ 2" is a dummy valued 1 if the subject is assigned to a group where the count of Strongly Altruistic types is equal or greater than 2. Indivdual controls include: age, gender, Econ major, the Big-Five personalities dimensions and a measure of numeracy.

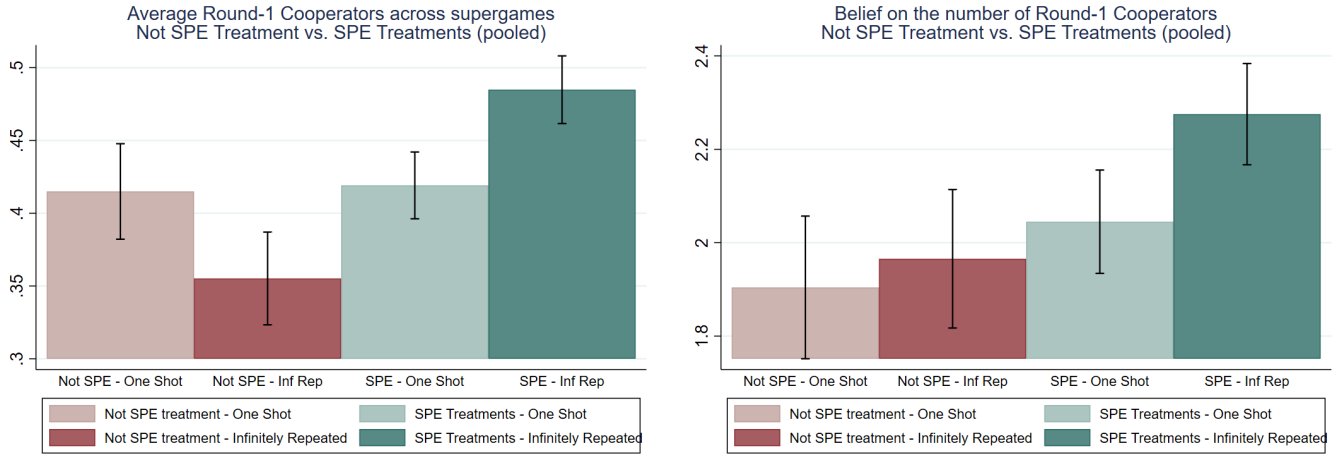
Clustering at matching group level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Result 4. *Strongly Altruistic types do not react differently than others to changes in strategic incentives for cooperation in terms of behavior, although they seem to react less strongly in terms of beliefs.*

Result 4 is supported by the evidence brought by Table 17 and Figures 12 and 13, which show average cooperation across supergames (Left panel) and average beliefs on the number of Round-1 cooperators in the group (Right panel), separately for One Shot and Infinitely Repeated play: in

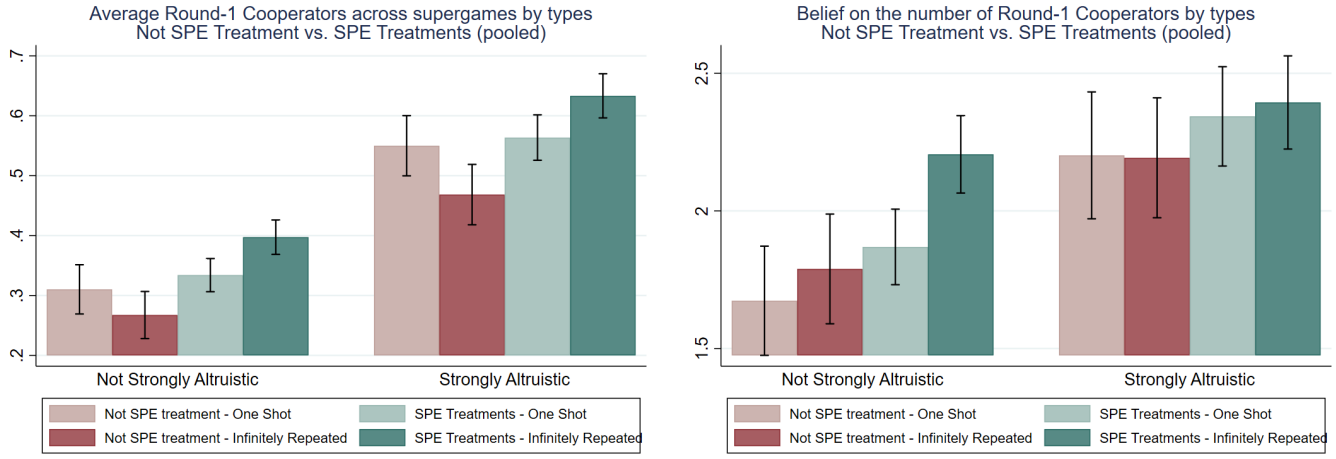
Figure 12, we distinguish observations from groups who have been exposed to treatments where cooperation is and is not sustainable as an equilibrium (SPE v. Not SPE); in Figure 12, we further break down the evidence by social preference type.

Figure 12: Changes in Round-1 Cooperation and Beliefs: One Shot vs. Infinitely Repeated play



Notes. Data from One Shot and Infinitely Repeated play (N=265); Legend: Not SPE treatment = Treatment 1 where $\delta = 0$ in the One Shot block and $\delta = 0.35$ in the Infinitely Repeated block (N=87). SPE treatments = Treatment 2 and 3, where where $\delta = 0$ in the One Shot block and $\delta = 0.55$ and $\delta = 0.75$, respectively, in the Infinitely Repeated block (N=178).

Figure 13: Changes in Round-1 Cooperation and Beliefs: One Shot vs. Infinitely Repeated play by types



Notes. Data from One Shot and Infinitely Repeated play (N=265); Legend: Not SPE treatment = Treatment 1 where $\delta = 0$ in the One Shot block and $\delta = 0.35$ in the Infinitely Repeated block (N=87). SPE treatments = Treatment 2 and 3, where where $\delta = 0$ in the One Shot block and $\delta = 0.55$ and $\delta = 0.75$, respectively, in the Infinitely Repeated block (N=178).

In Table 17, in order to estimate and test the relevance of the effect of changes in strategic incentives for cooperation on behavior, we adopt a Difference-in-Difference approach: we identify individuals

assigned to Treatment 1 - where $\delta = 0.35$ in the Infinitely Repeated block - as the Control group and individuals assigned to Treatment 2 and 3 - where $\delta = 0.55$ and $\delta = 0.75$ in the Infinitely Repeated block - as the Treated group. Since all individuals are exposed to a block of One Shot PDs (which corresponds to the pre-treatment period in our D-i-D framework, see Table 16), we want to test whether the change in cooperativeness moving from One Shot to Infinitely Repeated blocks is positive and significant for individuals in the Treated group, who are exposed to a significant change in strategic incentives for cooperation:

$$Y_{igt} = Y_0 + \alpha_g \cdot Treat + \gamma_t \cdot IR + \theta_* \cdot (Treat \cdot IR) + \epsilon_{igt}$$

where Y_{igt} is the binary cooperation choice by individual "i" belonging to group "g" (either Treated or Control) at time "t" (either belonging to the One Shot OS or infinitely repeated IR block) and the treatment effect is captured by parameter θ_*.

Table 16: Difference-in-Difference Framework

	Treated Group (T2 and T3)	**Control Group (T1)**
Pre-Treatment Period (One Shot play)	One Shot play T2 and T3	One Shot play T1
Post-Treatment Period (Infinitely Repeated play)	Inf. Rep. play T2 and T3	Inf. Rep. play T1

Table 17 shows the results of this estimation, proving there's a positive and significant treatment effect, as suggested by the graphical evidence brought by Figure 12 (Left panel). In order to test whether individuals belonging to different social preference types react differently to the treatment, we estimate a fully interacted model, in order to obtain an estimation of all paramenters of interest, and mainly of θ_*^{SA} and θ_*^{nSA}:

$$\begin{cases} Y_{igt}^{SA} = Y_0^{SA} + \alpha^{SA} \cdot Treat_g + \gamma^{SA} \cdot IR_t + \theta_*^{SA} \cdot (Treat_g \cdot IR_t) + \epsilon_{igt}^{SA} \\ Y_{igt}^{nSA} = Y_0^{nSA} + \alpha^{nSA} \cdot Treat_g + \gamma^{nSA} \cdot IR_t + \theta_*^{nSA} \cdot (Treat_g \cdot IR_t) + \epsilon_{igt}^{nSA} \end{cases}$$

Table 17 shows also the results of the estimation of the fully interacted model, which confirm what emerges from the graphical evidence reported in Figure 13 (Left panel). Strongly Altruistic types

have a significantly higher intercept (the hypothesis that $Y_0^{SA} = Y_0^{nSA}$ is strongly rejected in all specifications) and, as expected, there's no significant "group-assignment effect" for either social preference type (the hypothesis that $\alpha^{SA} = \alpha^{nSA}$ is accepted in all specifications). Similarly to what emerged from the pooled D-i-D model, there's a negative and significant "time-effect" when moving from One Shot to Infinitely Repeated blocks: the effect is not statistically different between the two social preference types (the hypothesis that $\gamma^{SA} = \gamma^{nSA}$ is accepted in all specifications) and could be explained by the fact that individuals belonging to the Control group de-facto experience a prolonged series of interactions in an environemnt where cooperation is not sustainable as an equilibrium and further learn not to cooperate.

The "treatment effect" is positive and significant in all specifications in which we do not control for individuals' beliefs and the effect is not statistically different between the two social preference types (the hypothesis that $\theta_*^{SA} = \theta_*^{nSA}$ is accepted in all specifications). Once we control for individual beliefs on the share of cooperators in the group, however, the effect vanishes. A cadidate explanation for this evidence, combined with what emerges from Figures 12 and 13 (Right panel) on the dynamics of beliefs across treatments and social preference types, is that not Strongly Altrustic types do react more strongly than Strongly Altruistic types to changes in strategic incentives for cooperation but only in terms of beliefs. However, the translation of this effect into behavior is weakened by the fact that Strongly Altruistic types have substantially higher perceptions on the share of cooperators in the group, regardless of the strategic environment. This further evidence of the dramatic role of beliefs is in line with what recently found by Aoyagi et al. [2020] on the role of beliefs in finitely and infinitely repeated games, and could also reconcile our evidence with the results found by Kölle et al. [2020] on Infinitely Repeated PDs with segregated groups, where beliefs are primed by design and polarize in the two sub-populations of the 'prosocial' and the 'selfish' [37].

[37] Kölle et al. [2020] sort their subjects based on their willingness to cooperate, as second movers, in a One Shot sequential PD, knowing that the first mover decided to cooperate: prosocial and selfish players are then segregated and informed of what was the choice of the other members of their group in the sequential PD before playing a series of Infinitely Repeated PDs. They find that groups formed only by prosocial players, where players are aware that they are playing with individuals of their same type - reach steadily higher levels of cooperation compared to groups of selfish players.

Table 17: Difference-in-Difference: Analysis of Changes in Round-1 Cooperation

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Round-1 Cooperation				Beliefs	
SPE	0.00624	-0.0208	0.0224	-0.00991	0.150	0.186
	(0.0414)	(0.0331)	(0.0386)	(0.0317)	(0.123)	(0.120)
Inf. Rep. (IR)	-0.0598*	-0.0708**	-0.0461	-0.0654	0.0613	0.107
	(0.0326)	(0.0279)	(0.0474)	(0.0413)	(0.121)	(0.167)
SPE x IR	0.126***	0.0950***	0.111**	0.0714	0.169	0.232
	(0.0407)	(0.0349)	(0.0555)	(0.0473)	(0.148)	(0.191)
Belief		0.181***		0.174***		
		(0.0112)		(0.0113)		
Strongly Altruistic (SA)			0.225***	0.136***		0.508***
			(0.0392)	(0.0328)		(0.114)
IR x SA			-0.0312	-0.0114		-0.106
			(0.0590)	(0.0521)		(0.208)
SPE x IR x SA			0.0337	0.0646		-0.190
			(0.0744)	(0.0659)		(0.249)
Constant	0.604*	0.117	0.394	-0.00930	2.693***	2.319***
	(0.310)	(0.249)	(0.297)	(0.246)	(0.833)	(0.811)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R-squared (overall)	0.0418	0.375	0.138	0.417	0.0366	0.0777
N	530	530	530	530	530	530

Notes: Models (1)-(6): Estimated coefficients of GLS Random Effect models. Models (1)-(4): Dependent variable = Average Round-1 Cooperation across supergames. Models (5)-(6): Average Belief on Round-1 Cooperators in the group across supergames.

"SPE" is a dummy valued 1 if the subject is exposed to a treatment where cooperation is sustainable as an equilibrium. "Inf. Rep." (IR) is a dummy valued 1 for observations regarding subjects' play in the Infinitely Repeated block. "Strongly Altruistic" (SA) is a dummy valued 1 if the subject is classified as a Strongly Altruistic type. "Belief" is the average belief on the number of Round-1 cooperators in the group. Indivdual controls include: age, gender, Econ major, the Big-Five personalities dimensions and a measure of numeracy.

Clustering at individual level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

7 Conclusion

We investigated what shapes individuals' cooperative attitudes in Prisoner Dilemmas.

From the meta-analysis it emerged that environmental incentives for cooperation, measured using compact indicators developed by the literature like $(\delta - \delta_{RD})$ or $sizeBAD$, do have some predictive power: their ability to predict Round-1 Cooperation choices increases over supergames, suggesting individuals actually learn to react to environmental incentives. It further emerged an asymmetry in terms of prediction accuracy along the SPE and RD equilibrium dimensions: ML algorithms seem to produce classifiers that are more accurate - especially in terms of models' ability to detect cooperators - in treatments where cooperation is sustainable in equilibrium, leaving a large fraction of unexplained variation, instead, in treatments where cooperation is not an equilibrium.

From the experimental data, collected through an online procedure that allows us to observe, within subjects, both individuals' social preferences' type and actual play in Infinitely Repeated PDs, we found a strong evidence in favor of the role of social preferences. Contrary to our initial conjecture, but not in contrast with our theoretical framework, social preferences affect cooperative attitude both when cooperation is and is not sustainable as an equilibrium. Moreover, we document that beliefs have a dramatic role in shaping cooperation and that individuals who exhibit strong social peferences have significantly higher beliefs on others' willingness to cooperate, which makes them less reactive in terms of beliefs to changes to strategic incentives for cooperation. The difference in beliefs between social preference types is mitigated when cooperation is an equilibrium and all individuals have on average optimistic perceptions on others' willingness to cooperate: accordingly, the concentration of individuals with strong social preferences affects cooperation levels in the former case, possibly through a beliefs' updating mechanism, but not in the latter.
Therefore, social preferences prove to be highly relevant in shaping cooperative attitudes both in One Shot and Infinitely Repeated PDs. Social preferences can thus represent a valuable tool to uncover some of the unexplained variation in cooperation observed so far, especially in contexts where cooperation is not an equilibrium and neither the standard theory nor the most recent experimental advances would provide any guidance to predict cooperation attainment.

References

Andreoni, J. and Miller., J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal*, 103(418):570–85.

Aoyagi, M., Fréchette, G. R., and Yuksel, S. (2020). Beliefs in repeated games.

Aoyagi, M. and Freéchette., G. R. (2009). Collusion as public monitoring becomes noisy: Experimental evidence. *Journal of Economic Theory*, 144(3):1135–65.

Arechar, A., Kouchaki, M., and Rand, D. (2018). Examining spillovers between long and short repeated prisoner's dilemma games played in the laboratory. *Games*, 9(1):5.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 255(6324):483–485.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4):1472–1503.

Blanco, M., Engelmann, D., and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

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

Blonski, M. and Spagnolo, G. (2015). Prisoners' other dilemma. *International Journal of Game Theory*, 44(1):61–81.

Bruhin, A., Fehr, E., and Schunk, D. (2018). The many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences. *Journal of the European Economic Association*, 17(4):1025–1069.

Bruttel, L. and Kamecke, U. (2012). Infinity in the lab. how do people play repeated games? *Theory and Decision*, 72(2):205–19.

Butler, J. V., Giuliano, P., and Guiso, L. (2015). Trust, values, and false consensus. *International Economic Review*, 56(3):889–915.

Capraro, V., Jordan, J. J., and Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments. *Scientific reports*, 4:6790.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131:141–150.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. (1996). Russell cooper and douglas v. dejong and robert forsythe and thomas w. ross. *Games and Economic Behavior*, 12(2):187–218.

Dal Bó, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review*, 95(5):1591–604.

Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5):2205–29.

Dal Bó, P. and Fréchette, G. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29.

Dal Bó, P. and Fréchette, G. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114.

Dal Bó, P. and Fréchette, G. (2019). Strategy choice in the infinitely repeated prisoners dilemma. *American Economic Review*. (forthcoming).

Davis, D., Ivanov, A., and Korenok, O. (2016). Individual characteristics and behavior in repeated games: An experimental study. *Experimental Economics*, 19(1):67–99.

Dreber, A., Fudenberg, D., and Rand., D. G. (2014). The role of altruism, inequality aversion, and demo- graphics. *Journal of Economic Behavior and Organization*, 98:41–55.

Dreber, A., Fudenberg, D., Rand., D. G., and Nowak, M. A. (2008). Winners don't punish. *Nature*, 452(7185):348–51.

Duffy, J. and Muñoz-García, F. (2012). Patience or fairness? analyzing social preferences in repeated games. *Games*, 3(1):56–77.

Duffy, J. and Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, 66(2):785–812.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fréchette, G. and Yuksel, S. (2017). Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination. *Experimental Economics*, 20(2):279–308.

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*. (forthcoming).

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–54.

Fudenberg, D. and Peysakhovich, A. (2016). Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)*, 4(4):23.

Fudenberg, D., Rand, D. G., and Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2):720–49.

Ghidoni, R., Cleave, B. L., and Suetens, S. (2019). Perfect and imperfect strangers in social dilemmas. *European Economic Review*, 1(116):148–59.

Greiner, B. (2004). The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, 10(23):63–104.

Harsanyi, J. C., Selten, R., et al. (1988). A general theory of equilibrium selection in games. *MIT Press Books*, 1.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *New York: Springer*.

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

KaterinaSherstyuk, Tarui, N., and Saijo., T. (2013). Payment schemes in in nite-horizon experimental games. *Experimental Economics*, 16(1):125–53.

Kim, J. (2019). The effects of time preferences on cooperation: Experimental evidence from infinitely repeated games. *Working Paper*.

Kölle, F., Quercia, S., and Tripodi, E. (2020). Social preferences under the shadow of the future. *Available at SSRN 3622125*.

Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252.

Murnighan, J. K. and Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games: A test of several models. *Journal of Con ict Resolution*, 27(2):279–300.

Naecker, J. (2015). The lives of others: Predicting donations with non-choice responses. *Working Paper*.

Naecker, J. and Peysakhovich, A. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior and Organization*, 133:373–384.

Nay, J. J. and Vorobeychik, Y. (2016). Predicting human cooperation. *PLoS ONE*, 11(5).

Peysakhovich, A., Nowak, M. A., and Rand, D. G. (2014). Humans display a 'cooperative phenotype'that is domain general and temporally stable. *Nature communications*, 5:4939.

Peysakhovich, A. and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647.

Proto, E., Rustichini, A., and Sofianos., A. (2019). Intelligence, personality, and gains from cooperation in repeated interactions. *Journal of Political Economy*, 127(3):1351–1390.

Reuben, E. and Suetens, S. (2012). Revisiting strategic versus non-strategic cooperation. *Experimental Economics*, 15(1):24–43.

Romero, J. and Rosokha, Y. (2018). The evolution of cooperation: The role of costly strategy adjustments. *American Economic Journal: Microeconomics*, 11(1):299–328.

Roth, A. E. and Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, 17(2):189–98.

Sabater-Grande, G. and Georgantzis., N. (2002). Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior and Organization*, 48(1):37–50.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., and Peters., E. (2013). Development and testing of an abbreviated numeracy scale: A rasch analysis approach. *Journal of Behavioral Decision Making*, 26(2):198–212.

Y. Zhao, Y. and Cen, Y. (2014). Data mining applications with r. *Oxford Academic Press: Elsevier*.

8 Appendix

A.1 Meta-Analysis: Dataset

In their paper, Dal Bó and Fréchette [2018] perform a regression analysis to test the ability of the two indices $\delta - \delta_{RD}$ and $sizeBAD$ to describe the levels of cooperativeness observed in their data by relying on a simple Probit model: they study the effect of the two indices on 1st Round Cooperation, separately looking at 1st Round Cooperation in Supergames 7 and 15 (see Table A1.1, a replication of Table 8 in Dal Bó and Fréchette [2018]).

Table A1.1: Round-1 Cooperation Cooperation - Marginal Effects at the average (DF2018)

	(1)	(2)	(3)	(4)	(5)	(6)
	Sup.7	Sup.15	Sup.7	Sup.15	Sup.7	Sup.15
SPE	-0.0986	0.195				
	(0.145)	(0.136)				
$(\delta - \delta_{SPE})$ x SPE	0.747***	0.979***				
	(0.0780)	(0.0733)				
$(\delta - \delta_{SPE})$ x Not SPE	0.566**	-0.349				
	(0.282)	(0.275)				
RD			0.113**	0.121***	0.225	0.420*
			(0.0451)	(0.0415)	(0.220)	(0.243)
$(\delta - \delta_{RD})$ x RD			1.030***	1.677***		
			(0.129)	(0.178)		
$(\delta - \delta_{RD})$ x Not RD			0.238***	0.235		
			(0.0574)	(0.273)		
$sizeBAD$ x RD					-0.902***	-1.139***
					(0.326)	(0.372)
$sizeBAD$ x Not RD					-0.429*	-0.342
					(0.229)	(0.368)
Observations	2,305	1,030	2,305	1,030	2,305	1,030

Notes: Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

We replicate the same analysis on our metadata, to show that the qualitatively results on the two indices on 1st Round Cooperation are unchanged (see Table A1.2).

Table A1.2: Round-1 Cooperation - Marginal Effects at the average (Metadata)

	(1)	(2)	(3)	(4)	(5)	(6)
	Sup.7	Sup.15	Sup.7	Sup.15	Sup.7	Sup.15
SPE	-0.0212	0.218**				
	(0.116)	(0.103)				
$(\delta - \delta_{SPE})$ x SPE	0.591***	0.723***				
	(0.121)	(0.156)				
$(\delta - \delta_{SPE})$ x Not SPE	0.519**	-0.350				
	(0.235)	(0.240)				
RD			0.134**	0.142***	0.240	0.325*
			(0.0536)	(0.0384)	(0.195)	(0.185)
$(\delta - \delta_{RD})$ x RD			0.844***	1.038***		
			(0.0999)	(0.182)		
$(\delta - \delta_{RD})$ x Not RD			0.249***	0.305		
			(0.0589)	(0.197)		
$sizeBAD$ x RD					-1.085***	-1.143***
					(0.253)	(0.173)
$sizeBAD$ x Not RD					-0.463**	-0.428*
					(0.217)	(0.250)
Observations	3,157	1,616	3,157	1,616	3,157	1,616

Notes: Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Table A1.3: General information on the meta-data

	Session	**Subjects**	δ	g	l	Supergames
Andreoni and Miller (1993)	**1**	**14**	0	1.67	1.33	200
Cooper et al. (1996)	**3**	**33**	0	0.44	0.78	10
Dal Bó (2005)	**6**	**276**				
	2	72	0	1.17	0.83	7
	2	102	0	0.83	1.17	9
	1	42	0.75	1.17	0.83	7
	1	60	0.75	0.83	1.17	10
Dreber et al. (2008)	**2**	**50**				
	1	28	0.75	2	2	21
	1	22	0.75	1	1	27
Aoyagi and Fréchette (2009)	**4**	**74**				
	2	36	0	0.33	0.11	75
	2	38	0.9	0.33	0.11	10
Duffy and Ochs (2009)	**9**	**102**	0.9	1	1	13
Dal Bó, Foster and Putterman (2010)	**28**	**424**	0	1	3	10
Dal Bó and Fréchette (2011)	**18**	**266**				
	3	44	0.5	2.57	1.86	71
	3	50	0.5	0.67	0.87	72
	3	46	0.5	0.09	0.57	77
	3	44	0.75	2.57	1.86	33
	3	38	0.75	0.67	0.86	47
	3	44	0.75	0.09	0.57	35
Blonski, Ockenfels and Spagnolo (2011)	**10**	**200**				
	1	20	0.5	2	2	11
	2	40	0.75	2	2	11
	1	20	0.75	1	8	11
	1	20	0.75	1	1	11
	1	20	0.75	0.83	0.5	11
	1	20	0.75	0.75	1.25	11
	1	20	0.75	0.5	3.5	11

	1	20	0.875	2	2	8
	1	20	0.875	0.5	3.5	8
Fudenberg, Rand and Dreber (2012)	**3**	**48**	0.875	0.33	0.33	9
Bruttel and Kamecke (2012)	**3**	**36**	0.8	1.17	0.83	2
Sherstyuk, Tarui, and Saijo (2013)	**4**	**56**	0.75	1	0.25	29
Frechette and Yuksel (2017)	**3**	**60**	0.75	0.4	0.4	12
Dal Bó and Fréchette (2019)	**41**	**672**				
	3	50	0.5	2.57	1.86	37
	8	140	0.5	0.09	0.57	46
	8	114	0.75	2.57	1.86	25
	10	164	0.75	0.09	0.57	24
	10	168	0.9	2.57	1.86	21
	2	36	0.95	2.57	1.86	7
Peysakhovich and Rand (2016)	**6**	**96**				
	3	52	0.125	0.66	0.33	45
	3	44	0.875	0.33	0.33	10
Romero and Rosokha (2018)	**6**	**82**				
	3	44	0.95	2.57	1.86	10
	3	38	0.95	2.57	1.86	20
Ghidoni, Cleave and Suetens (2019)	**4**	**80**	0	0.73	0.46	10
Proto, Rustichini and Sofianos (2019)	**40**	**586**				
	32	476	0.75	0.09	0.57	12
	8	110	0.5	0.09	0.57	13
	201	**3267**	Choices: 269.832			

Notes. The column 'Supergame' reports the minimum number of supergames observed across similar sessions.

A.1.1 The Logistic LASSO

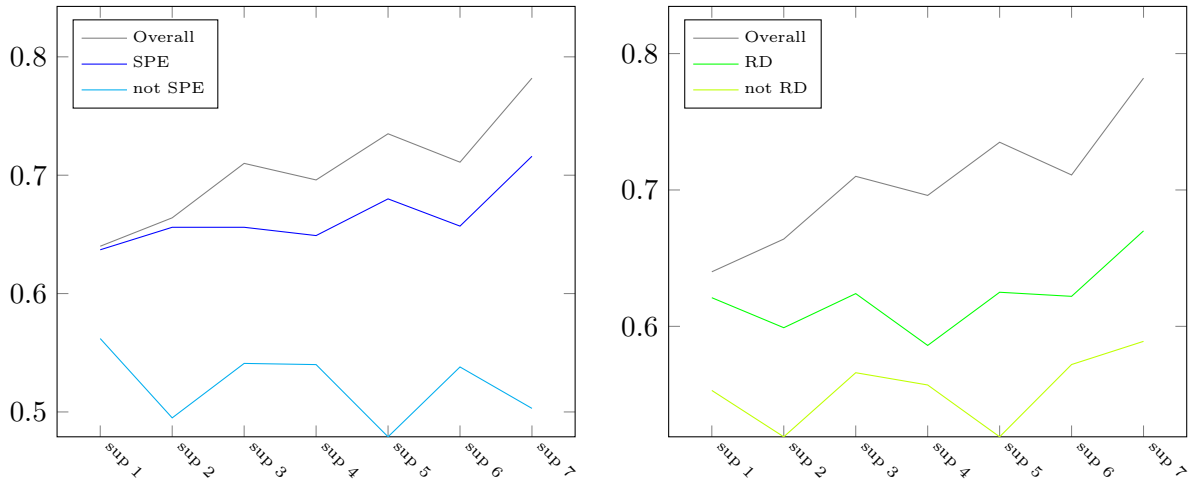
The logistic LASSO algorithm provides a prediction that is based on a logit model (with a linear index) where the estimated coefficients are penalized according to their magnitude:

$$max_{\beta_0,\beta} \sum_{i=1}^{N} [y_{i,t}(\beta_0 + \beta' \tilde{X_{i,t}}) - log(1 + e^{\beta_0 + \beta' \tilde{X_{i,t}}})] - \sum_{j=1}^{\tilde{P}} |\beta_j|$$

where $\tilde{X_{i,t}}$ (of dimension $\tilde{P}$) is the vector of all predictors including also the pairwise interactions between the variables in S_t, λ is a penalization parameter [38], β_0 is a constant and β' is a transpose vector of the β coefficients to be estimated (together with the constant) [39]. The LASSO penalization implies that only a subset of indicators will have estimated coefficients other than zero and the non-null coefficients of the sparse model will be the only ones relevant for prediction.

A.2 Meta-Analysis: Additional Figures

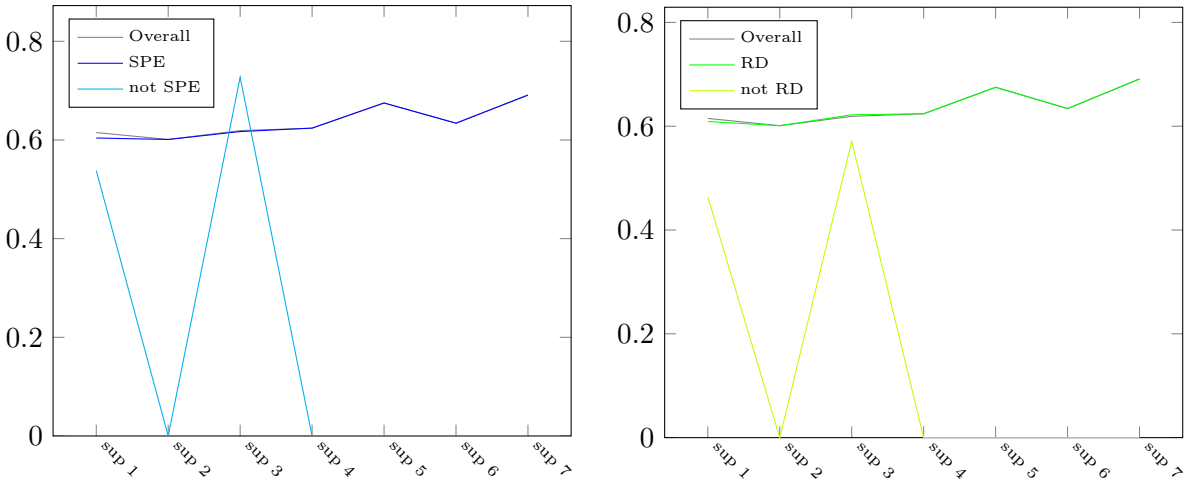
Figure A2.1: Classification Accuracy metrics over Supergames 1 to 7 (*sizeBAD* model)



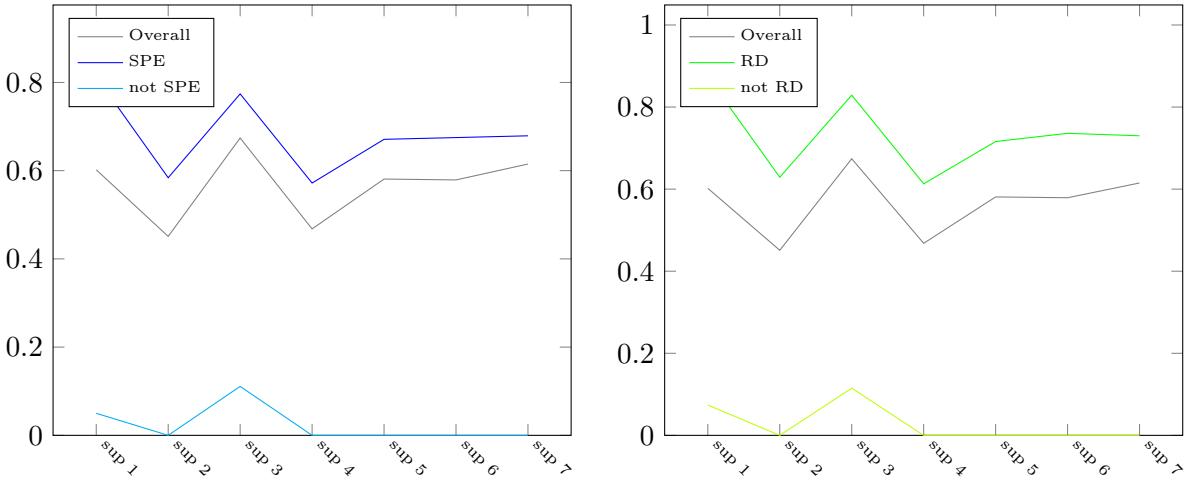
(a) Area under the ROC curve: Supergame 1 to 7

[38] The optimal λ is selected by looking at the 10-fold cross-validated misclassification error, using the one-standard-error rule. The cross-validation process aims at identifying the value of λ that minimizes the misclassification error, but acknowledging that the estimation of misclassification rates also comes with an error, we adopt the "one-standard error" rule: according to this rule, we choose the most parsimonious model whose error is no more than one standard error above the error of the best model, which represents a more conservative approach (Hastie et al. [2009]).

[39] The index $i = 1, .., N$ identifies each single individual in the sample, who is exposed to one treatment $t = 1, .., T$ only. We observe cooperation realizations at the individual level ($y_{i,t}$, such that $y_{i,t} \neq y_{l,t}$), but we feed our algorithm with treatment-specific variables only ($\tilde{X_t} = \tilde{X_{i,t}} = \tilde{X_{l,t}}$). The set of regressors in $\tilde{X_t}$ changes depending of which of the three models, see Table 2, we are estimating $\tilde{X_t^k}$, where $k = \{1, 2, 3\}$.

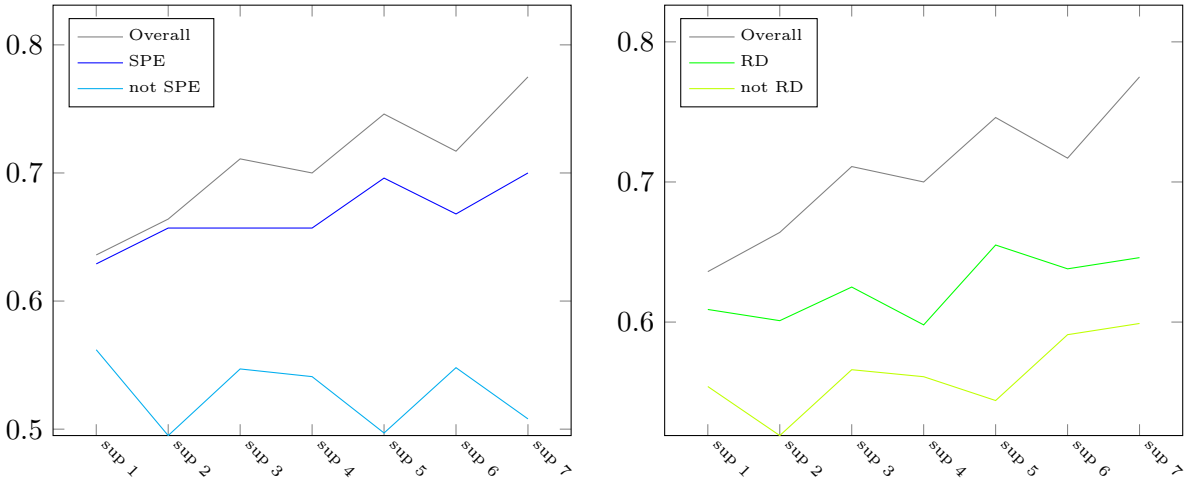


(b) Precision: Supergame 1 to 7

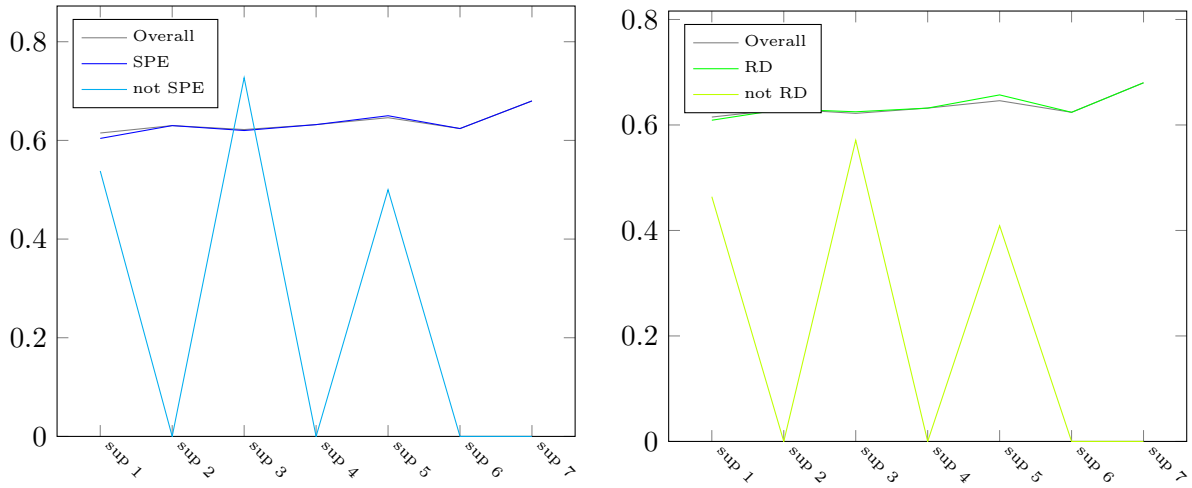


(c) Recall: Supergame 1 to 7

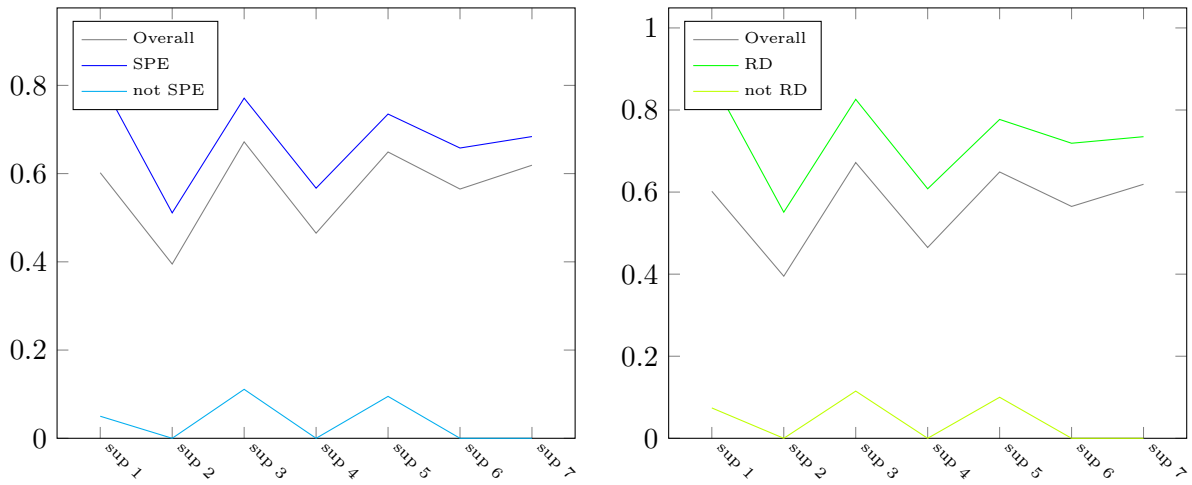
Figure A2.2: Classification Accuracy metrics over Supergames 1 to 7 (Unconstrained model)



(a) Area under the ROC curve: Supergame 1 to 7



(b) Precision: Supergame 1 to 7



(c) Recall: Supergame 1 to 7

A.3 Theoretical Framework: Modeling social preferences à la Fehr and Schmidt [1999]

When we model social preferences using the original model by Fehr and Schmidt [1999] both concerns for advantageous and disadvantageous inequality are considered. In this case, focusing on a two-players case with complete information, the utility function of individual i in the pair would be given by:

$$U_i(A_i, A_j) = f(\pi_i(A_i, A_j), \pi_j(A_i, A_j)) = \pi_i - a_i max\{\pi_j - \pi_i; 0\} - b_i max\{\pi_i - \pi_j; 0\}$$

such that the utility of individual i can also be expressed as:

$$\begin{cases} if\ \pi_i = \pi_j \longrightarrow U_i = \pi_i \\ if\ \pi_i > \pi_j \longrightarrow U_i = \pi_i - b_i(\pi_i - \pi_j) \\ if\ \pi_i < \pi_j \longrightarrow U_i = \pi_i - a_i(\pi_j - \pi_i) \end{cases}$$

In this framework, a_i and b_i measure weights attached to differences in payoffs within the pair both in situations of advantageous and disadvantageous inequality, where it is usually further assumed that $a_i \geq b_i$ and $1 > b_i \geq 0$, which implies that the weight assigned to the envy-driven dis-utility from having the lowest payoff in the pair is higher than the weight assigned to the guilt-driven dis-utility from having the highest payoff in the pair.

Using a different utility function to map payoffs into utilities alters the reformulation of the stage-game matrix in terms of players' utilities shown in 9, as shown in A3.1.

Table A3.1: Prisoners' Dilemma Row Player's Utilities - F&S

General			Social Preferences à la Fehr and Schmidt (1999)		
	C	D		C	D
C	$U_i(C,C)$	$U_i(C,D)$	C	$U_i(C,C) = R$	$U_i(C,D) = S - a_i(T-S)$
D	$U_i(D,C)$	$U_i(D,D)$	D	$U_i(D,C) = T - b_i(T-S)$	$U_i(C,C) = P$

In this context, if players have perfect information about social preferences, as shown by Duffy and Muñoz-García [2012], we have that for no or low levels of social preferences ($b_i = 0$ or $b_i \leq b_i^* = \frac{(T-R)}{(T-S)}$), the only Nash Equilibrium (NE) of the stage game if (Defect, Defect), but if both players have strong enough social preferences ($b_i > b_i^* = \frac{(T-R)}{(T-S)}$), then multiple NE - (Cooperate,Cooperate), (Defect, Defect) and a mixed strategy equilibrium - emerge. Accordingly, once we move to the infinitely repeated version of the game, the presence of social preferences allows cooperation to arise as a:

- Subgame Perfect Nash equilibrium,
  whenever $\delta_i > \delta_i^{SPE} = \frac{(T-R)-b_i(T-S)}{(T-P)-b_i(T-S)}$
  where $\delta_i^{SPE} : \sum_{t=0}^{\infty} \delta^t R > T - b_i(T-S) + \sum_{t=1}^{\infty} \delta^t P$

- Risk Dominant equilibrium,

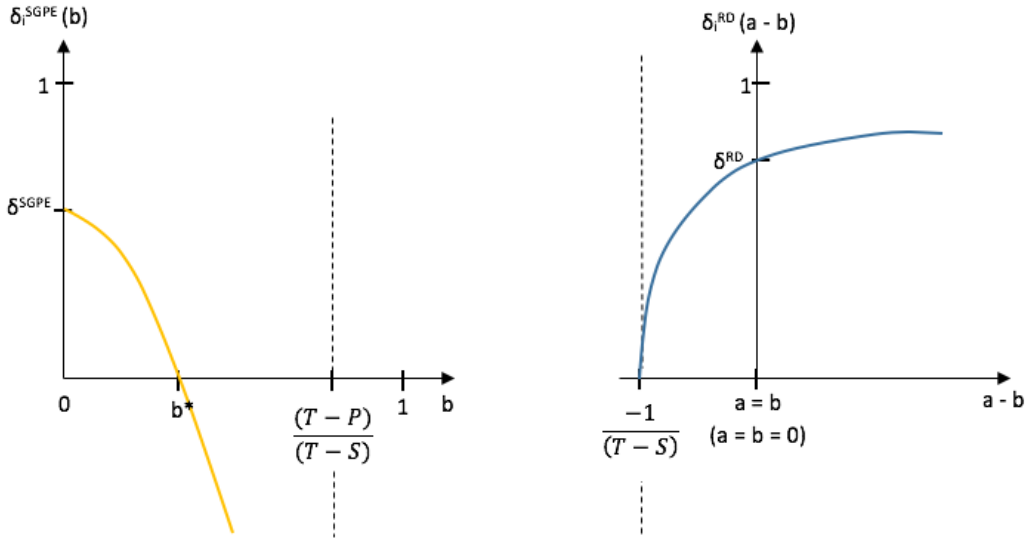
  whenever $\delta_i > \delta_i^{RD} = \frac{(P-R)+(T-S)(1+a_i-b_i)}{(T-S)(1+a_i-b_i)}$

  where, δ_i^{RD} :

  $\frac{1}{2}\left[\sum_{t=0}^{\infty}\delta^t R\right]+\frac{1}{2}\left[S-a_i(T-S)+\sum_{t=1}^{\infty}\delta^t P\right] > \frac{1}{2}\left[T-b_i(T-S)+\sum_{t=1}^{\infty}\delta^t P\right]+\frac{1}{2}\left[\sum_{t=0}^{\infty}\delta^t P\right]$

In both cases, where $a_i = b_i = 0$, δ_i^{SPE} and δ_i^{RD} coincide with threshold values δ^{SPE} and δ^{RD}, while for high enough values of b_i, $\delta_i^{SPE} < \delta^{SPE}$ and $\delta_i^{RD} < \delta^{RD}$ for $a_i - b_i < 0$ and $\delta_i^{RD} > \delta^{RD}$ for $a_i - b_i > 0$, being both δ_i^{SPE} and δ_i^{RD} decreasing in b_i.

Figure A3.1: δ_i^{SPE} and δ_i^{RD} as a function of b_i and $a_i - b_i$



A.4 Experimental Design: Measuring social preferences

A.4.1 Comparing results with and w/o reciprocity in Bruhin et al. [2018]: Discussion

We evaluate how sensitive the results of the paper are to the inclusion/exclusion of reciprocity concerns in the social preferences model. We compare the estimates reported in the paper, obtained on the whole sample, which includes observations from both the Dictator Games (DG) and the Reciprocity Games (RG) for all individuals, with estimates obtained estimating the constrained model (where $\gamma = 0$ and $\eta = 0$) only on the observations collected from the Dictator Game.

$$U_i = (1 - \alpha s - \beta r - \gamma q - \eta v) \cdot \Pi^i + (\alpha s + \beta r + \gamma q + \eta v) \cdot \Pi^j$$

where:

$s = 1$ if $\Pi^i < \Pi^j$, and $s = 0$ otherwise (disadvantageous inequality);

$r = 1$ if $\Pi^i > \Pi^j$, and $s = 0$ otherwise (advantageous inequality);

$q = 1$ if player j behaved kindly toward i, and $q = 0$ otherwise (positive reciprocity);

$v = 1$ if player j behaved unkindly toward i, and $v = 0$ otherwise (negative reciprocity);

In the first case, we rely on all the information from the 117 binary decisions taken from the subjects throughout the experiment (117 x 174 = 20358 observations) , in the second case, we restrict our attention to the 39 binary DG decisions (39 x 174 = 6786 observations). We evaluate whether:

A) the estimates of the parameters derived from the aggregate model (n. types K=1) are substantially different

B) the summmary statistics of the individual estimates of the parameters (n. types K=174) are substantially different

C the estimates of the parameters derived from the finite mixture model (n. types K=3) are substantially stable across types

We run this analysis both for observations from Session 1 and Session 2 of the Experiment. We further investigate, for estimates obtained from Session 2 data, whether:

D) the ability to explain the variability in subjects' subsequent choices in the Trust Games (TG) and the Reward and Punishment Games (RPG) using linear models augmented with predictions based on finite-mixture and inidividual model estimates, is substantially different

In the paper, a McFadden's (1981) random utility model for discrete choices is used to estimate the social preference parameters of the behavioral model $\theta = (\alpha,\ \beta,\ \gamma,\ \eta)$. The underlying assumption is that the utility player i gets from choosing the allocation $X_g = (\Pi^i_{X_g},\ \Pi^j_{X_g},\ r_{X_g},\ s_{X_g},\ q_{X_g},\ v_{X_g})$ in game $g = 1, .., G$, within the set of possible allocations $\{X_g,\ Y_g\}$ is given by:

$$U^i(X_g;\theta,\sigma) = U^i(X_g;\theta) + \epsilon_{X_g}$$

where $U^i(X_g;\theta)$ is the deterministic component of the utility deriving from allocation X_g and ϵ_{X_g} is a random component representing noise in the utility evaluation: the random component ϵ_{X_g} is assumed to follow a type 1 extreme value distribution with a scale parameter $\frac{1}{\sigma}$. Under this framework, player i would choose allocation X_g over the allocation Y_g whenever $U^i(X_g;\theta,\sigma) \geq U^i(Y_g;\theta,\sigma)$, so that the probability that the choice of player i in game g - C_g - equals X_g is given by:

$$Pr(C_g = X_g;\ \theta,\ \sigma,\ X_g,\ Y_g)$$

$$= Pr(U^i(X_g;\theta) - U^i(Y_g;\theta) \geq \epsilon_{Y_g} - \epsilon_{X_g}$$

$$= \frac{exp(\sigma U^i(X_g;\theta))}{exp(\sigma U^i(X_g;\theta)) + exp(\sigma U^i(Y_g;\theta))}$$

where the parameter σ measures choice sensitivity to differences in deterministic utilities, so that when $\sigma = 0$ player i chooses each of the two options with the same 0.5 probability irrespective of the deterministic utility associated to the two options, while when σ is arbitrarily large the probability of choosing the most appealing option in terms of deterministic utility approaches 1.

The first approach estimates the random utility model by pooling the data to obtain aggregate estimates of the parameters $(\hat{\theta},\ \hat{\sigma})$. These aggregate estimates represent the most parsimonious characterization of social preferences, where all players are assumed to belong to the same 'representative' type (n. types = 1).

At the opposite extreme, the individual estimates are obtained by separately estimating the parameters of the social preferences model for each individual $(\hat{\theta_i},\ \hat{\sigma_i})$. This approach is the least parsimonious, and likely to suffer from small sample bias, but is able to fully uncover the behavioral heterogeneity in the data (n. types = N).

The intermediate approach in terms of flexibility and parsimony is represented by the finite mixture model, where the population is assumed to be characterized by a finite number of K distinct preference types, each characterized by a different set of parameters $(\hat{\theta_k},\ \hat{\sigma_k})$. This approach acknowledges latent heterogeneity in the data, although individual type-membership is not directly observable. In

this context, the estimation leads to a parsimonious characterization of the K types in the population, providing a set of type-specific preference parameters and types' shares in the population $\hat{\pi_k}$. In the paper, the optimal number of types is fixed to K=3.

Models reported in the following pages are estimated on observations from the full set of 174 players. For individual estimates, summary statistics are reported on the sample of 160 players whose estimated parameters are not classified as erratic, based on the estimates obtained in the paper.

A.4.2 Comparing results with and w/o reciprocity in Bruhin et al. [2018]: Data Session 1

A) Estimates from the aggregate model

DG Sample		Paper Sample	
$\hat{\alpha}$	0.0628***	$\hat{\alpha}$	0.0835***
	(0.016)		(0.015)
$\hat{\beta}$	0.279***	$\hat{\beta}$	0.261***
	(0.021)		(0.019)
$\hat{\sigma}$	0.016***	$\hat{\sigma}$	0.016***
	(0.001)		(0.001)
		$\hat{\gamma}$	0.072***
			(0.014)
		$\hat{\eta}$	-0.042***
			(0.011)

B) Summary of individual estimates

DG Sample					Paper Sample				
	MIN	MAX	MED	MEAN		MIN	MAX	MED	MEAN
$\hat{\alpha_i}$	-13.783	0.459	0.052	-0.101	$\hat{\alpha_i}$	-1.394	0.471	0.053	0.017
$\hat{\beta_i}$	-11.336	1.085	0.197	0.1669	$\hat{\beta_i}$	-1.977	0.998	0.211	0.216
$\hat{\gamma_i}$	-	-	-	-	$\hat{\gamma_i}$	-0.366	0.783	0.042	0.0836
$\hat{\eta_i}$	-	-	-	-	$\hat{\eta_i}$	-1.106	0.598	-0.008	0.055
$\hat{\sigma_i}$	0.000	0.804	0.0599	0.2915	$\hat{\sigma_i}$	0.004	0.858	0.035	0.174

C) Estimates from the Finite Mixture model

DG Sample				Paper Sample			
	BA	MA	SA		BA	MA	SA
$\hat{\pi_k}$	0.179***	0.352***	0.469***	$\hat{\pi_k}$	0.121***	0.474***	0.405***
$\hat{\alpha_k}$	-0.38***	0.044***	0.149***	$\hat{\alpha_k}$	-0.435***	0.065***	0.159***
$\hat{\beta_k}$	-0.016	0.077***	0.482***	$\hat{\beta_k}$	-0.145	0.129***	0.463***
$\hat{\gamma_k}$	-	-	-	$\hat{\gamma_k}$	0.17	-0.001	0.151***
$\hat{\eta_k}$	-	-	-	$\hat{\eta_k}$	-0.076	-0.027**	-0.053***
$\hat{\sigma_k}$	0.009***	0.066***	0.020***	$\hat{\sigma_k}$	0.008***	0.0316***	0.018***

Using estimated posterior probabilities to belong to each type, we get the same classification for $15 + 49 + 55 = 119/160 = 74\%$ of the subjects.

Paper Sample	**DG Sample**			
	BA	MA	SA	Total
BA	15	4	0	19
MA	10	49	17	76
SA	4	6	55	65
Total	29	59	72	160

Session 2

A) Estimates from the aggregate model

DG Sample		Paper Sample	
$\hat{\alpha}$	0.079***	$\hat{\alpha}$	0.098***
	(0.013)		(0.013)
$\hat{\beta}$	0.255***	$\hat{\beta}$	0.245***
	(0.020)		(0.019)
$\hat{\sigma}$	0.020***	$\hat{\sigma}$	0.019***
	(0.001)		(0.001)
		$\hat{\gamma}$	0.029***
			(0.010)
		$\hat{\eta}$	-0.043***
			(0.008)

B) Summary of individual estimates

DG Sample					Paper Sample				
	MIN	MAX	MED	MEAN		MIN	MAX	MED	MEAN
$\hat{\alpha}_i$	-1.240	0.449	0.052	0.035	$\hat{\alpha}_i$	-1.636	0.399	0.060	0.048
$\hat{\beta}_i$	-0.362	1.012	0.160	0.232	$\hat{\beta}_i$	-0.405	0.905	0.169	0.225
$\hat{\gamma}_i$	-	-	-	-	$\hat{\gamma}_i$	-1.087	0.679	0.005	0.032
$\hat{\eta}_i$	-	-	-	-	$\hat{\eta}_i$	-0.0553	0.229	-0.009	0.045
$\hat{\sigma}_i$	0.007	0.929	0.471	0.389	$\hat{\sigma}_i$	0.007	0.886	0.069	0.275

C) Estimates from the Finite Mixture model

DG Sample				Paper Sample			
	BA	MA	SA		BA	MA	SA
$\hat{\pi}_k$	0.102***	0.449***	0.449***	$\hat{\pi}_k$	0.100***	0.544***	0.356***
$\hat{\alpha}_k$	-0.368***	0.042***	0.160***	$\hat{\alpha}_k$	-0.328***	0.061***	0.193***
$\hat{\beta}_k$	-0.047	0.072***	0.469***	$\hat{\beta}_k$	-0.048	0.095***	0.495***
$\hat{\gamma}_k$	-	-	-	$\hat{\gamma}_k$	-0.028	-0.005	0.099***
$\hat{\eta}_k$	-	-	-	$\hat{\eta}_k$	-0.015	-0.019***	-0.082***
$\hat{\sigma}_k$	0.018***	0.085***	0.023***	$\hat{\sigma}_k$	0.015***	0.049***	0.019***

Using estimated posterior probabilities to belong to each type, we get the same classification for $15 + 67 + 56 = 138/160 = 86\%$ of the subjects.

Paper Sample	**DG Sample**			
	BA	MA	SA	Total
BA	15	0	1	16
MA	1	67	19	87
SA	0	1	56	57
Total	16	68	76	160

Following the same approach as in the paper, we further analyze whether linear models augmented with predictions based on finite-mixture and inidividual model estimates can better explain the variability in the choice made by the same set of subjects in a series of Trust Games (TG) and Reward and Punishment Games (RPG). The models are augmented with, respectively:

- Type-specific prediction on the probability to take the choice of interest
- Individual-specific prediction on the probability to take the choice of interest
- Type-specific prediction on the probability to take the choice of interest *and* Difference between the Individual-specific prediction and the Type-specific prediction (Δ_{i-p}).

The augmented models are compared to a baseline model estimated using only individual-specific characteristics, such as Big 5 personality traits, cognitive ability, age, gender, monthly income, and field of study as explanatory variables.

DG Sample				**Paper Sample**			
Trust Game							
Indiv. charact.	$\rightarrow$	$R^2 = 0.0589$	-	Indiv. charact.	$\rightarrow$	$R^2 = 0.0589$	-
Indiv. charact.	$\rightarrow$	$R^2 = 0.3655$	$\hat{\beta}_{tp} = 0.617^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.3491$	$\hat{\beta}_{tp} = 0.607^{***}$
+ Type pred.				+ Type pred.			
Indiv. charact.	$\rightarrow$	$R^2 = 0.3677$	$\hat{\beta}_{ip} = 0.606^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.3457$	$\hat{\beta}_{ip} = 0.580^{***}$
+ Indiv. pred.				+ Indiv. pred.			
Indiv. charact.	$\rightarrow$	$R^2 = 0.4040$	$\hat{\beta}_{tp} = 0.686^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.3748$	$\hat{\beta}_{tp} = 0.650^{***}$
+ Indiv. pred.			$\hat{\beta}_{\Delta} = 0.344^{***}$	+ Indiv. pred.			$\hat{\beta}_{\Delta} = 0.309^{***}$
+ Δ_{i-t}				+ Δ_{i-t}			

DG Sample				**Paper Sample**			
Reward-Punishment Game							
Indiv. charact.	$\rightarrow$	$R^2 = 0.0354$	-	Indiv. charact.	$\rightarrow$	$R^2 = 0.0354$	-
Indiv. charact.	$\rightarrow$	$R^2 = 0.2127$	$\hat{\beta}_{tp} = 0.971^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.2674$	$\hat{\beta}_{tp} = 1.123^{***}$
+ Type pred.				+ Type pred.			
Indiv. charact.	$\rightarrow$	$R^2 = 0.194$	$\hat{\beta}_{ip} = 0.532^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.253$	$\hat{\beta}_{ip} = 0.641^{***}$
+ Indiv. pred.				+ Indiv. pred.			
Indiv. charact.	$\rightarrow$	$R^2 = 0.229$	$\hat{\beta}_{tp} = 0.898^{***}$	Indiv. charact.	$\rightarrow$	$R^2 = 0.302$	$\hat{\beta}_{tp} = 1.064^{***}$
+ Indiv. pred.			$\hat{\beta}_{\Delta} = -.254^{***}$	+ Indiv. pred.			$\hat{\beta}_{\Delta} = -.353^{***}$
+ Δ_{i-t}				+ Δ_{i-t}			

A.4.3 Bruhin et al. [2018] Design: Dictator Games

DG	Π_X^i	Π_X^j	Π_Y^i	Π_Y^j
1	940	150	800	510
2	970	490	770	170
3	1060	330	680	330
4	990	480	750	180
5	930	510	810	150
6	430	1030	230	710
7	370	1060	290	680
8	350	1060	310	680
9	1010	190	730	470
10	420	1040	240	700
11	450	1020	210	720
12	470	730	190	1010
13	870	140	870	520
14	400	690	260	1050
15	350	680	310	1060
16	950	510	790	150
17	910	520	830	140
18	390	1050	270	690
19	330	680	330	1060
20	890	140	850	520
21	410	1050	250	690
22	1050	270	690	390
23	520	870	140	870
24	890	520	850	140
25	510	810	150	930
26	960	500	780	160
27	620	790	580	410
28	670	420	530	780
29	720	750	480	450
30	700	760	500	440
31	680	780	520	420
32	740	460	460	740
33	620	410	580	790
34	790	600	410	600
35	660	780	540	420
36	690	770	510	430
37	600	410	600	790
38	640	790	560	410
39	780	540	420	660

A.5 Experimental Design: Recruitment & Matching Procedures

We recruit participants in order to have 16 subjects per session in Part 2 of the experiment [40].

When subjects take part in Part 1, playing the DGs in the role of the dictator, they are informed their actions will have monetary consequences both on their payoff and on the payoff of their matched partner, who is a randomly selected individual from their same session, with whom they will never interact again in Part 2 of the Experiment.

When subjects take part in Part 2, and interact in real time to play the infinitely repeated PDs, they are grouped in 4 groups of 4 people ($N_G = 4$) before each of the two δ-blocks starts. Subjects are informed that they will never interact with the same counterpart across the two different δ-blocks, and that not even their counterparts from the two different δ-block will ever interact with each other. The matching structure, as shown in Figure A5.2, ensures that subject i:

- will never meet again his/her counterparts from the δ'-block in the δ''-block;
- within each δ-block, will be randomly paired with any of his/her $N_G - 1 = 3$ group mates, with a probability to be rematched to the same partner from the previous round equal to $1/3$;
- in Part 2, will never interact again with the partners he/she was matched to in Part 1 (subject i' counterpart in Part 1 is randomly selected from the pool of the $N-1-((N_G-1)*2) = 15-6 = 9$ subjects not grouped with subject i in Part 2).

[40] To this aim, we recruit a higher number of participants in order to hit the target of least 20-22 subjects completing Part 1 within the due date and being entitled to participate in Part 2. This allows us to manage attrittion issues between Part 1 and Part 2. Redundant participants who complete Part 1 within the due date and log-in on time for Part 2, but do not actually play, are paid a show-up fee of 5 Euros.

Figure A5.2: Matching procedure for subject $A1$ in Part 1 and Part 2

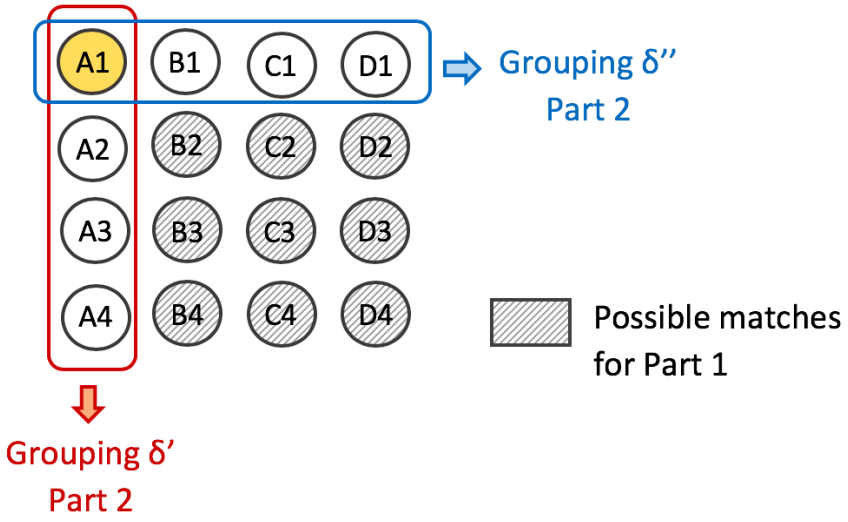
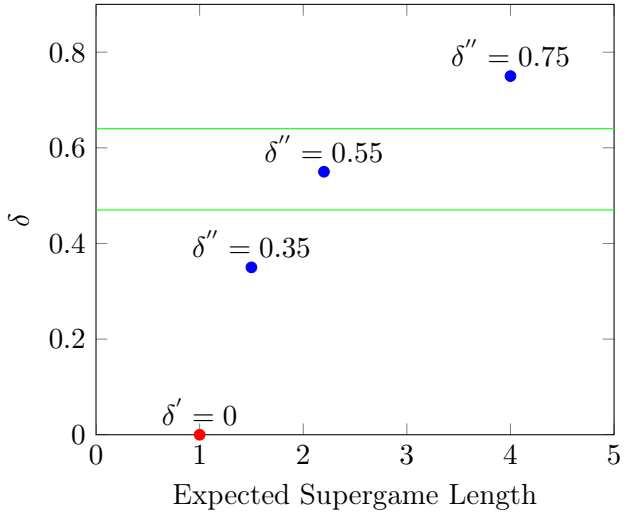


Figure A5.1: Part 2 - Differences across Treatments



	$\delta - \delta_{SPE}$	$\delta - \delta_{RD}$	$sizeBAD$
$\delta' = 0$	-0.47	-0.64	1
$\delta'' = 0.35$	-0.12	-0.29	1
$\delta'' = 0.55$	+0.08	-0.09	0.736
$\delta'' = 0.75$	+0.28	+0.11	0.3

A.6 Experimental Design: Instructions

Translation of the instructions from Italian.

A.6.1 Part 1

Welcome.

Thanks for your participation in this study!

During this study, you will have the opportunity to earn money. The amount of your earnings will depend on the decisions you and the other participants will make. All decisions will remain completely anonymous.

The study is divided into two parts: a preliminary part and the main part of the study.

Before moving to the main part of the study, you are required to take part in the preliminary part of the study, which will last about 30 minutes. You can take part in the preliminary part of the study whenever you prefer before $*date\_time\_ddl\_part1*$. Remember that you will be paid for your participation in the study ONLY IF you complete the preliminary part before the deadline and log in on time for the main part of the study. Payments will only be calculated at the end of the main part of the study.

This preliminary part of the study is divided into two parts: Part A and Part B. At the beginning of each part you will get the corresponding instructions and we will ask you to answer a few comprehension questions.

- In part A we will ask you to decide how certain monetary payments between you (Person "A") and another specific participant in the experiment (Person "B") should be distributed.
- In Part B we will ask you to complete a questionnaire.

Part A: Instructions

In this part of the study you will have to make 39 decisions that will affect you and another participant, who will be randomly selected among the other participants in this study and will be paired with you in each decision situation. You will never learn who this person is, and the other person

will also not learn of your identity. You will no longer interact with this participant for the rest of the study.

In each of the 39 decisions you will have exactly two options: X and Y.

Each option is associated with a monetary amount for you (Player A) and for the other participant paired with you (Player B). With your decision you will determine the distribution of payments between you and the other participant definitively, the other participant has a passive role and cannot change the distribution.

Please note: We present monetary amounts as points on the computer screen. 100 points are worth 0.4 Euros.

Payments: You will be paid only for one of the decisions you will make, which will be randomly selected at the end of the entire study. At the end of the study, you will be informed of which decision has been randomly selected for payment and of how much you and the other participant will receive, based on your choices.

	Amount for You (A)	**Amount for Participant B**	Your choice
Distribution X	1040	600	X
Distribution Y	850	850	Y

The 39 different situations will be presented successively on the screen, like in this example, where the payments associated with each of the two options - X and Y - are shown for you and the other participant: in this case, if you choose X you will receive 1040 points and the other participant 600 points, while if you choose Y you both receive 850 points.

Before we start with Part A, we will ask you to answer some comprehension questions.

[Part A - Control Questions]

Part B: Instructions

This part consists of a questionnaire. It is important for us that you answer the questions as good as possible.

(1) Demographic data [Year of birth | Gender | Major]

(2) 44-items Big Five Inventory (John and Srivastava [1999])

(3) 8-items Numeracy test (Weller et al. [2013])

Thank you for your time!
You will receive further instructions and the link to log in for the main part of study in the next days by email. Remember that you will be entitled to receive the show up fee of 5 Euros only if you log in on time to take part in the main part of the study.

A.6.2 Part 2

Welcome.
Thanks for your participation in this study.
By logging-in on time, after successfully completing the preliminary part of the study, you have earned the right to receive a payment of minimum 5 Euros.
All the decisions you will make during the study will remain completely anonymous. The final amount of your earnings will depend on the decisions you and the other participants will make during the study.
At the end of today's session, we will inform you of the amount of your earnings based on the decisions that you and the other study participants will make today and made in the preliminary part of the study.

Please shut down all the other programs running on your computer except Zoom, which you should keep open until the very end of the study.
All you will need is a blank sheet of paper. It is important that you do not try to communicate with

the other participants during the session.

Today we will ask you to take part in two activities [41].: in both activities you will have to make a series of decisions but at the end of the study you will be actually paid only for one of the decisions you have madein each activity , which will be selected randomly at the end of the study.

Please note: We present monetary amounts as points on the computer screen. 100 points are worth 1.5 Euros.

Activity n.1: Instructions (One Shot block)

Task 1 consists of 10 rounds. In each round you will interact with a counterpart and you will be asked to make a decision.

Before the first round starts, you will be paired with three other participants to this study, who will be the members of your group for the entire duration of activity n.1. At the beginning of each interaction round, you will be paired with another participant, randomly selected among your group mates.

You will not learn the identity of the participant paired with you and he will not learn about yours. Over the ten rounds you may be re-paired with a participant with whom you have already interacted in one of the previous rounds but you will not be able to identify when this may happen.

In each round, you and the other participant will have two possible choices: X and Y. Each cell shows the amount of your earnings in points (left, in blue) and that of the other participant (right, in black).

		The other participant	
		X	**Y**
You	**X**	73 pts ; 73 points	10 pts ; 100 points
	Y	100 pts ; 10 points	43 pts ; 43 points

[41] The order of Activities n. 1 and n. 2 was randomized across sessions.

Before rounds 1, 5 and 10 begin, we will also ask you to guess what number of participants in your group who will choose the 'X' option. If, at the end of the study, one of these rounds is randomly selected for payment and your conjecture proves correct, you will receive an additional fixed payment of 2 Euros.

Before we start with Activity n.1, we will ask you to answer some comprehension questions.

[Activity n.1 - Control Questions]

Activity n.2: Instructions (Infinitely Repeated block)

Activity 2 consists of 10 matches. Each match consists of a variable number of rounds.

Before the first match starts, you will be paired with three other participants to this study, who will be the members of your group for the entire duration of activity n.2. None of these participants interacted with you during activity n.1.

Each match can consist of one or more rounds of interaction. After each round of each match, we will randomly draw a number within the interval [1,100].

> If the number drawn is <= [δ : *continuation probability*], the match continues for another round.
>
> If the drawn number is> [δ : *continuation probability*], the match ends.

The duration of each match is therefore determined randomly and there is a probability of [δ : *continuation probability*]% that the match continues for another round.

At the beginning of each match, you will be paired with a partner, randomly selected among your group mates. You will interact with the same partner for the entire duration of the match. Over the ten matches you may be re-paired with a participant with whom you have already interacted in one of the previous matches but you will not be able to identify when this may happen.

In each round of each match, you and the other participant will have two possible choices: X and Y.

Each cell shows the amount of your earnings in points (left, in blue) and that of the other participant (right, in black).

		The other participant	
		X	Y
You	X	73 pts ; 73 points	10 pts ; 100 points
	Y	100 pts ; 10 points	43 pts ; 43 points

You will only be paid for the decisions you will make in one of the matches, which will be randomly selected at the end of the study. Your earnings will be equal to your overall earnings, which correspond to the sum of the earnings you have realized through all the rounds of the selected match.
Before matches 1, 5 and 10 begin, we will also ask you to guess what number of participants in your group who will choose the option 'X' in the first round of that match. If, at the end of the study, one of these matches is randomly selected for payment and your conjecture on the first round proves correct, you will receive an additional fixed payment of 2 Euros.
Before we start with Activity n.2, we will ask you to answer some comprehension questions.

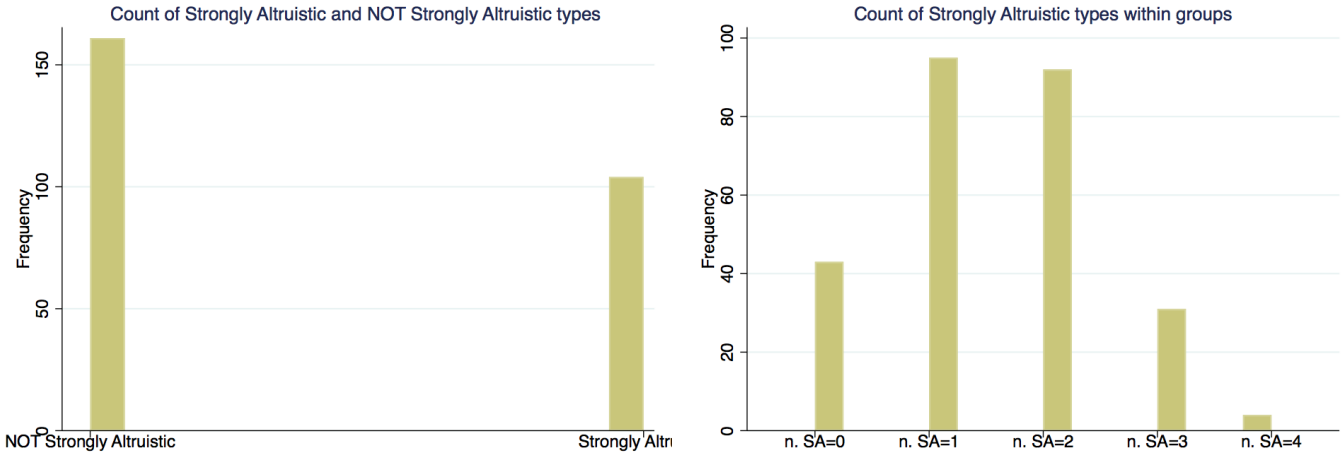
[Activity n.2 - Control Questions]

A.7 Results: Additional descriptive statistics

A.7.1 Sample: Descriptive Statistics

Figure A7.1 shows the distribution of social preference types within the sample (Left panel) and the, endogenously determined, distribution based on the number of Strongly Altruistic types within the matching group (Right panel).

Figure A7.1: Distribution by social preference type and SA count (N=265)



Notes. Data from all sessions with $\delta = 0$ (N=265); Legend. SA: 'Strongly Altruistic'; not SA: 'not Strongly Altruistic', which combines 'Behindness-Averse' and 'Moderately Altruistic' individuals.

A.7.2 Mean Round1-Cooperation over supergames by social preferences types

Table A7.1: Mean of Round1-Cooperation over supergames - One Shot play

	SA = 0 T1	SA = 0 T2	SA = 0 T3	SA = 1 T1	SA = 1 T2	SA = 1 T3
	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)
Sup: 1	.5510204	.5636364	.5263158	.7631579	.7777778	.9666667
	.5025445	.5005048	.5037454	.4308515	.421637	.1825742
	49	55	57	38	36	30
Sup: 2	.4693878	.5272727	.4385965	.6315789	.6944444	.9333333
	.5042338	.5038572	.5006262	.4888515	.4671766	.2537081
	49	55	57	38	36	30
Sup: 3	.244898	.4181818	.3508772	.7368421	.6111111	.7
	.434483	.4978066	.4814868	.4462583	.4944132	.4660916
	49	55	57	38	36	30
Sup: 4	.2653061	.3454545	.3333333	.4473684	.5	.6333333
	.4460713	.479899	.4755949	.5038966	.5070926	.4901325
	49	55	57	38	36	30
Sup: 5	.3673469	.5090909	.3157895	.6842105	.5833333	.7666667
	.4870779	.504525	.4689614	.4710691	.5	.4301831
	49	55	57	38	36	30
Sup: 6	.2653061	.4545455	.1578947	.5526316	.4722222	.5666667
	.4460713	.5025189	.3678836	.5038966	.5063094	.5040069
	49	55	57	38	36	30
Sup: 7	.2653061	.2545455	.1403509	.3947368	.4444444	.4333333
	.4460713	.4396203	.3504383	.4953554	.5039526	.5040069
	49	55	57	38	36	30
Sup: 8	.2653061	.2909091	.1052632	.4210526	.3888889	.4333333
	.4460713	.4583678	.3096202	.5003555	.4944132	.5040069
	49	55	57	38	36	30
Sup: 9	.1632653	.2727273	.122807	.4210526	.3333333	.3666667
	.3734378	.4494666	.3311331	.5003555	.4780914	.4901325
	49	55	57	38	36	30
Sup: 10	.244898	.4	.1754386	.4473684	.4166667	.3333333
	.434483	.4944132	.3837227	.5038966	.5	.4794633
	49	55	57	38	36	30
Total	.3102041	.4036364	.2666667	.55	.5222222	.6133333
	.4630498	.4910728	.4426051	.4981496	.5002011	.4877999
	490	550	570	380	360	300

Table A7.2: Mean of Round1-Cooperation over supergames - Infinitely Repeated play

	SA = 0 T1	SA = 0 T2	SA = 0 T3	SA = 1 T1	SA = 1 T2	SA = 1 T3
	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)
Sup: 1	.3877551	.5636364	.4035088	.6578947	.6388889	.7666667
	.4922875	.5005048	.4949621	.4807829	.4871361	.4301831
	49	55	57	38	36	30
Sup: 2	.3469388	.4909091	.3508772	.5526316	.5555556	.7666667
	.4809288	.504525	.4814868	.5038966	.5039526	.4301831
	49	55	57	38	36	30
Sup: 3	.244898	.4181818	.4561404	.5263158	.5277778	.8333333
	.434483	.4978066	.5025	.5060094	.5063094	.379049
	49	55	57	38	36	30
Sup: 4	.244898	.4363636	.4035088	.5	.4444444	.9
	.434483	.5005048	.4949621	.5067117	.5039526	.3051286
	49	55	57	38	36	30
Sup: 5	.2857143	.4727273	.5087719	.5526316	.5	.9
	.4564355	.5038572	.5043669	.5038966	.5070926	.3051286
	49	55	57	38	36	30
Sup: 6	.2040816	.3272727	.3859649	.4736842	.5	.7333333
	.4072055	.4735424	.4911497	.5060094	.5070926	.4497764
	49	55	57	38	36	30
Sup: 7	.2040816	.3636364	.3859649	.3157895	.5277778	.8333333
	.4072055	.4854794	.4911497	.4710691	.5063094	.379049
	49	55	57	38	36	30
Sup: 8	.2857143	.3454545	.3684211	.3947368	.5	.7333333
	.4564355	.479899	.4866643	.4953554	.5070926	.4497764
	49	55	57	38	36	30
Sup: 9	.1632653	.2727273	.3157895	.2631579	.3611111	.7666667
	.3734378	.4494666	.4689614	.4462583	.4871361	.4301831
	49	55	57	38	36	30
Sup: 10	.3061224	.2727273	.4035088	.4473684	.4444444	.7
	.4656573	.4494666	.4949621	.5038966	.5039526	.4660916
	49	55	57	38	36	30
Total	.2673469	.3963636	.3982456	.4684211	.5	.7933333
	.4430272	.4895869	.4899665	.4996597	.5006959	.4055908
	490	550	570	380	360	300

A.7.3 Mean Round1-Cooperation over supergames by social preferences types' concentration within groups

Table A7.3: Mean of Round1-Cooperation by SA types' concentration - One Shot play

	Count SA (0)	Count SA (1)	Count SA (2)	Count SA (3)	Count SA (4)
	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)
1	.1636364	.3965517	.4870968	.48125	.15
	.1747726	.2945732	.3480823	.301593	.057735
	11	29	31	16	4
2	.2611111	.4148148	.6066667	.525	
	.2226548	.2891682	.3084257	.34935	
	18	27	30	12	
3	.1357143	.3666667	.4645161	1	
	.1905746	.2688507	.333215	0	
	14	39	31	3	
Total	.1953488	.3894737	.5184783	.5483871	.15
	.2046524	.2803773	.3327968	.3365224	.057735
	43	95	92	31	4

Table A7.4: Mean of Round1-Cooperation by SA types' concentration - Infinitely Repeated play

	Count SA (0)	Count SA (1)	Count SA (2)	Count SA (3)	Count SA (4)
	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)	(mean/sd/N)
1	.2916667	.2666667	.38	.4583333	.55
	.1975225	.2140026	.358756	.4077841	.341565
	12	24	35	12	4
2	.4083333	.5606061	.3352941	.4166667	
	.3848455	.3749495	.2922041	.3298301	
	12	33	34	12	
3	.4181818	.4695652	.6058824	.7375	
	.3672762	.4353001	.3749272	.373927	
	22	23	34	8	
Total	.3826087	.44625	.4398058	.5125	.55
	.3342018	.3721112	.3606475	.3833216	.341565
	46	80	103	32	4