A Test of Loyalty

Renaud Foucart[*] Jonathan H.W. Tan[†]

June 27, 2023

Abstract

We propose and test a model of loyalty in games. Players can mutually maintain loyalty by working towards a common goal that is Pareto-superior to any Nash equilibrium without it. Loyalty imposes a a psychological cost on defecting in an ongoing cooperation, which is thus sustained. We distinguish loyalty from reciprocity and explain how it complements guilt aversion with two dynamic games from a field experiment conducted in a Pakistani factory. The evidence supports the validity of loyalty, which has a stronger effect within than between groups.

JEL codes: C72, C93, D91

Keywords: loyalty, cooperation, trust, reciprocity, field experiment

[*]Lancaster University Management School, r.foucart@lancaster.ac.uk

[†]Corresponding author: j.tan@ntu.edu.sg; Department of Economics, School of Social Sciences, Nanyang Technological University (NTU), Singapore. Acknowledgments: We are grateful to Bilal and Ali Riaz, Jamil Khan, Shamaz Khalid, Mustafa Zaman, and other assistants for supporting the experiment. Thanks to Friedel Bolle, Yves Breitmoser, Paul Fenn, Georg Kirschsteiger, Alexander Kritikos, Yohanes Eko Riyanto, Robert Sugden, Jiwei Zheng, and participants to talks in Nottingham and Lille for the advice and encouragement. Research funding under the NTU Start-Up Grant and MOE AcRF Tier 1 grant is gratefully acknowledged.

1 Introduction

The success of a group in achieving its goal rests on the mutual belief that its members will maintain cooperation and not defect midway for selfish gain. In this paper, we show that even payoff-irrelevant actions can create such beliefs, and help sustain cooperation. *Loyalty* describes the adherence of individuals to the goals, symbols, and beliefs of the group (James and Cropanzano, 1994). One's contribution to the group fosters the sense of loyalty (Zdaniuk and Levine, 2001). Social identity reinforces loyalty and motivates individuals to remain in a group to continue contributing rather than to leave midway for selfish gain (Van Vugt and Hart, 2004). Loyalty is therefore instrumental as a principle of partiality toward the group as an object (Hildreth et al., 2016). One's aversion to being disloyal increases when individuals are socially closer and more connected (Tang et al., 2017). Loyalty is found in alliances (Koza and Lewin, 2000), between friends (Ghobadian et al., 2007), coworkers (Sherony and Green, 2002), political partisans (Burgess, 1999), and sports club fans as consumers (Funk and James, 2006). Even nationalism originates from group loyalty (Druckman, 1994). Whereas pledges of loyalty can increase the salience of ethical behavior, it has a dark side that can also drive corruption (Hildreth et al., 2016). Despite being a ubiquitous social mechanism, the economics of loyalty remains unstudied.

We define loyalty and distinguish it theoretically and experimentally from other pro-social motives, thereby clarifying its role in strategic interactions. The challenge in disentangling loyalty from other motives is due to their confluence across various strategic contexts. For instance, "loyalty is represented in adherence to ingroup norms and trustworthiness in dealings with fellow ingroup members" (Brewer and Brown, 1998, p. 560). Further, surveys have evidenced the effect of group loyalty on social preferences by showing that it increases preference for redistribution (Luttmer, 2001). Thus, we identify loyalty by one's adherence to the group's goals, which is a characteristic of loyalty emphasized in the literature. Specifically, we model loyalty as a preference for maintaining a history of cooperation, built on beliefs that are mutually upheld when players take actions toward mutually beneficial outcomes. Unlike social preferences such as fairness (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) or efficiency (Engelmann and Strobel, 2004) concerns, the monetary payoffs of others do not directly enter an individual's psychological payoff. Unlike reciprocity, whereby one responds to their belief of others having kind intentions towards them (Dufwenberg and Kirchsteiger, 2004; Falk and

Fischbacher, 2006), loyalty requires neither risky investments of trust nor costly deviations from self-interested payoff maximization, whereas subsequent betrayal yields disutility in the form of treason costs. We see loyalty as complementary to guilt aversion. With guilt aversion, one bears a psychological cost for not fulfilling the expectations of others. Loyalty adds a component of beliefs on the intention of co-players to uphold a cooperative relationship. Loyalty rests on a history of perceived cooperation amongst players following observable actions that are consistent with it, whereas guilt can only be sustained by expectations. Loyalty is also consistent with the result that individuals tend to adhere to a norm of cooperative behavior, the "principle of mutual benefit", in voluntary interactions (Isoni et al., 2022).

We distinguish loyalty from other social preferences by comparing behavioral patterns between two games, each with two stages and two players, that yield identical payoff distributions across the respective outcomes. We introduce a novel *ROSCA game* (RG), inspired by the peer-lending microfinance arrangement commonly known as "committees" or Rotating Savings and Credit Associations (Anderson and Baland, 2002). Players receive an income at the start of each stage. In Stage 1, two players simultaneously choose whether to cooperate by joining the committee and contributing their stage incomes to the common "pot," which multiplies their income. If anyone refuses, then the game ends with players keeping their incomes from both stages. If both cooperate, then Nature randomly selects one of the two players to win the pot in Stage 1, modelling the benefit from the committee "loan" that allows one to advance a consumption costing more than the per stage income, e.g. repairing a broken roof that costs two stages of nominal income. In Stage 2, the winner from Stage 1 chooses whether to contribute her stage income to the pot, which the other player now wins. This contribution doubles in value for the winner, who would otherwise be left with only the Stage 2 income.[1]

We contrast this to a directly comparable adaptation of the *Trust Game* (Berg et al., 1995, TG) with similar payoffs and actions. However, in Stage 1, only the first mover can invest trust in the second mover by sending his Stage 1 income. Similar to the RG and also the trust game of Berg et al. (1995), the truster's investment triples in value for the trustee. If the first mover invests trust, the second mover then chooses whether or not to return money to the first mover. Like the RG, money returned by the second mover doubles in value for the

[1] The returned amount doubles in Stage 2, which is less than how it is tripled in Stage 1 to model the opportunity cost of further delayed consumption in Stage 2 as opposed to greater time-discounted benefits of receiving it in advance in Stage 1, as featured in real world settings.

first mover. This Pareto efficient outcome models the fulfilment of trust (Berg et al., 1995). If the second mover does not return, the first mover is left with his Stage 2 income and the second mover keeps the income from both stages and the tripled Stage 1 income sent by the first mover, which models the abuse of trust (Berg et al., 1995). For direct comparability of the TG and RG, subjects play the TG as first mover in one game and second mover in another, and are randomly paid for one of the two games.

The strategic difference between the two games is clearest from the perspective of reciprocity (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) and guilt aversion (Battigalli and Dufwenberg, 2009; Attanasi et al., 2016a; Di Bartolomeo et al., 2019).[2] With reciprocal preferences, the second mover reciprocates the "kind" trust shown by the first mover. Hence, there exists a reciprocal equilibrium in the TG in which the first mover is willing to trust if he expects the second mover to cooperate in return with sufficiently high probability. In the RG however, the decision to contribute in Stage 1 is not as kind as in TG (in the sense of Dufwenberg and Kirchsteiger, 2004), as sending in Stage 1 of the RG is compatible with self-interest. Reciprocity therefore predicts lower adherence to cooperation in the second stage of RG than in TG.

Guilt aversion allows for possible cooperation both in the TG and RG. As the two games feature identical payoffs for the various possible outcomes from the respective moves, the parameter values such that cooperating in Stage 2 is part of an equilibrium for a given set of beliefs are identical. The main difference however is that, by cooperating in the TG, the first mover signals his belief that the second mover will also cooperate. A second mover who did not expect cooperation initially therefore needs to update her second-order beliefs about the first mover's expectations upon observing his cooperation. In the RG, cooperating in the first stage is compatible with the expectation of no cooperation in Stage 2, so that no player needs to update their beliefs. Guilt aversion therefore also predicts lower adherence to cooperation in the second stage of the RG than in the TG, for given guilt aversion parameters and prior beliefs of cooperation in the second stage.

The joint initial stage in the RG enables the elicitation of loyalty. Both players make a first move, and form beliefs over each other's intent in Stage 2. We posit that this initial

[2] Outcome-based utilities also predict the same rates of sending and returning in both games, as returning in Stage 2 decreases inequality (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) or increases the social surplus (Engelmann and Strobel, 2004). Thus, these theories cannot distinguish between the two games.

move generates a sense of loyalty if both players expect that the intention of the co-player is to continue cooperating in the next stage. Such perceived cooperation introduces an additional loyalty cost in the Stage 2 of the RG but not in the TG. Thus, if loyalty preferences are sufficiently strong, players should cooperate more in Stage 2 in the RG than in TG. This result stands in contrast with reciprocity and guilt aversion. Observing a higher rate of Stage 2 cooperation in RG would therefore validate loyalty as a distinct preference, albeit one that should not substitute but complement other social preferences. Our main test is on the validity of loyalty preferences, such as theoretically modeled in conjunction with guilt aversion here, which is not the only way to model loyalty. Indeed, while introducing a loyalty cost in psychological utility seems to us the most natural way to represent this concept, we could derive similar predictions in a model where we represent loyalty as increasing guilt aversion and mutual beliefs in cooperation in the second stage.

The data was collected from a field experiment conducted in a textile factory in Lahore, Pakistan, which offers a relevant and representative sample as detailed in the experimental section 3. We observe that the majority of subjects contribute in Stage 1 and return in Stage 2 of the RG, and that more subjects return in the RG than in the TG. To test the sensitivity of loyalty to social identification, we compare loyalty effects within and between endogenously formed groups, which is also another feature of real world committees. It could increase cooperation through intrinsic social preferences for the ingroup (Chen and Li, 2009) or guilt aversion supported by the beliefs that others expect more cooperation from the ingroup (Charness et al., 2022; Ciccarone et al., 2020).[3] We validate group identification effects by showing that self-reported trust attitudes towards others within one's group are more favorable than towards others in the factory (i.e. between groups), which are in turn more favorable than others in general. Consistently, we also find that loyalty is more likely to emerge among subjects within these endogenously formed groups.

A large literature has focused on the role of reciprocity (Bacharach et al., 2007; Stanca et al., 2009) and guilt aversion in trust games (Guerra and Zizzo, 2004; Charness and Dufwen-

[3]For example, Ciccarone et al. (2020) distinguish between the two explanations by manipulating the second-order beliefs of dictators on the first-order beliefs of recipients on whether they are matched with an ingroup or outgroup dictator. Guilt aversion should generate a difference while social preferences should not, and they found evidence for the latter. However, Charness et al. (2022) found that group biases in guilt aversion can be reduced by making mutual promises that bridge the gap between members of different groups.

berg, 2006; Chang et al., 2011; Bracht and Regner, 2013; Ederer and Stremitzer, 2017; Bellemare et al., 2019) and modified trust games (Attanasi et al., 2019b). Our ROSCA game belongs to the latter category. Our paper also contributes to the literature focusing specifically on the respective roles of reciprocity and guilt aversion (Attanasi et al., 2013, 2019a), finding that guilt aversion is the most widespread belief-dependent motivation in trust games.

We present the theory in Section 2. Section 3 describes the experimental design and logistics. We report the results in Section 4. Section 5 discusses and concludes.

2 Theory

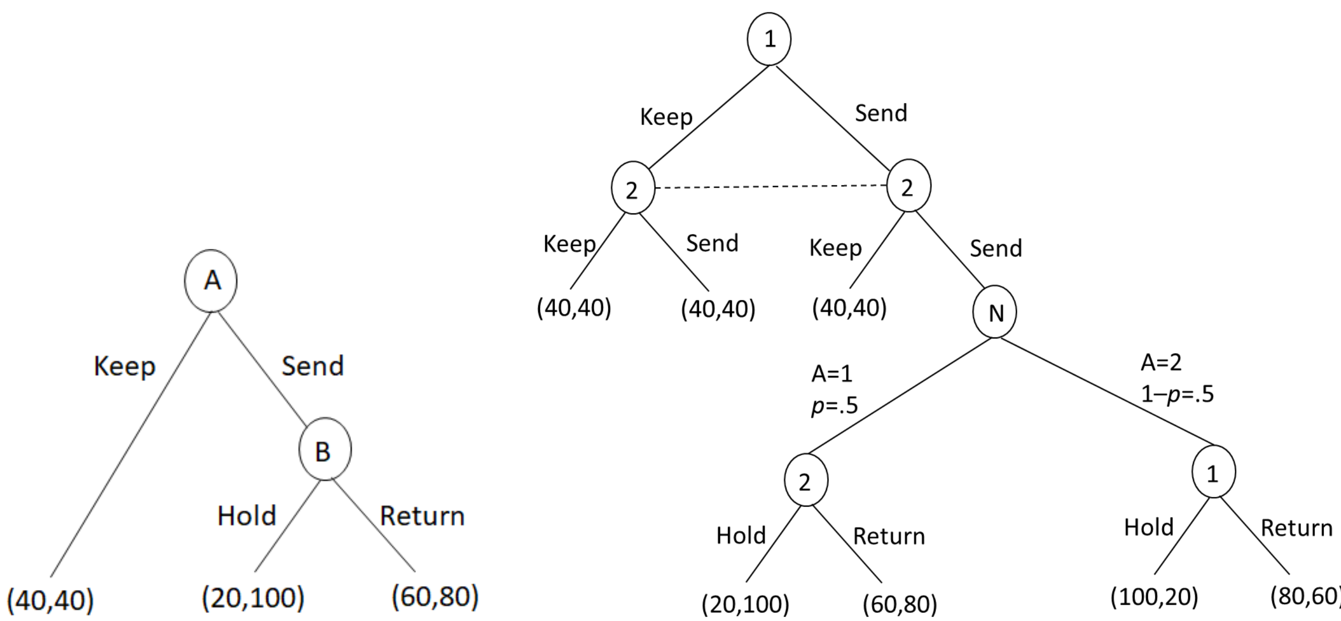
2.1 The Games

We discuss the predictions of existing theories for our two experimental games, starting with the assumption of self-interest.

Our first game, Γ_1, is the Trust Game (TG, Figure 1a). It is an adaptation of the game by Berg et al. (1995). Two risk-neutral players A (he) and B (she) both receive an initial endowment of 40. In Stage 1, player A chooses to KEEP or SEND part of his endowment. If he chooses KEEP, the game ends with payoffs $\{40, 40\}$, which is equal to the initial endowment (the first number in the payoff pair refers to the payoff of player A and the second to that of player B). If he chooses SEND, the value of the amount sent increases and player B must choose to HOLD or RETURN part of the amount received. If she chooses HOLD, their payoffs are $\{20, 100\}$, while if she chooses RETURN, their payoffs are $\{60, 80\}$. Assuming both players are only *self-interested*, the unique subgame perfect Nash equilibrium is (KEEP, HOLD).

Our second game, Γ_2, is the ROSCA Game (RG, Figure 1b). In Stage 1, players 1 (he) and 2 (she) simultaneously choose to SEND or KEEP a share of their endowment of 40. If at least one player chooses KEEP, similar to the TG the game ends and their monetary payoffs are $\{40, 40\}$ (the first number in the payoff pair refers to the payoff of player 1 and the second to that of player 2). We denote players in the RG by an index $\{1, 2\}$ in Stage 1 as they are symmetric prior to Nature's move. The game progresses to Stage 2 if both players SEND, upon which Nature randomly selects one of the players with probability $1/2$ to play a subgame similar to that of player B in the TG (i.e. $A = 1$ and $B = 2$ with $p = .5$, and $A = 2$ and $B = 1$ with $1 - p = .5$), while the other player who is assigned the role of A has no further choices to

Figure 1: Experimental games with material payoffs



(a) Γ_1 – Trust Game

(b) Γ_2 – ROSCA Game

make. The corresponding outcomes for A and B are similar to those in the TG.

With only self-interested players, there exists a subgame perfect Nash equilibrium in which both choose to SEND in Stage 1, and HOLD as player B. Indeed, in the last subgame, a player always receives a higher monetary payoff by choosing HOLD. By backward induction, player 1 expects that both he and player 2 HOLD with certainty when given a chance to do so. Hence, if a player believes that the other would SEND, it is a best response to SEND if

$$\frac{1}{2}100 + \frac{1}{2}20 > 40, \tag{1}$$

which always holds. An equilibrium in which both players choose KEEP in Stage 1 also exists, but it is not robust to a small tremble in the probability of the other player choosing KEEP.[4]

Prediction 0. With only self-interest, we would therefore expect subjects to never SEND and never RETURN in TG, and to always SEND and never RETURN in RG.

It is well known however that self-interest does a poor job at explaining observed behavior in the TG, thus we consider plausible reasons for players to RETURN: reciprocity and guilt aversion.

2.2 Reciprocity

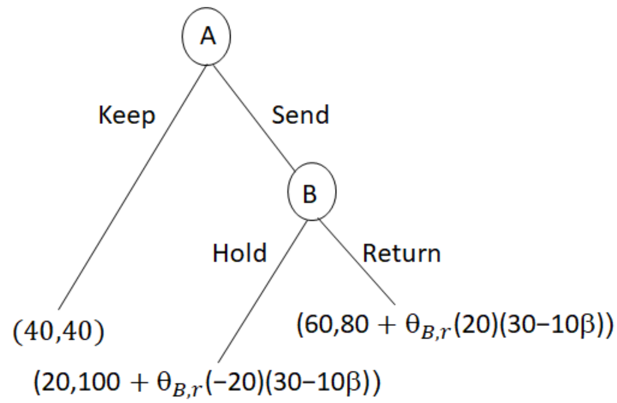
We first introduce reciprocity in the TG game described in Γ_1. Figure 2 presents Γ_3, the dynamic psychological TG (Battigalli and Dufwenberg, 2009) using the notion of sequential reciprocity of Dufwenberg and Kirchsteiger (2004). In that framework, the utility of the second mover B depends on a reciprocity component and takes the form

$$U_B = \pi_B + \theta_{Br} K_B \hat{\lambda}_B,$$

where π_B is the monetary payoff, θ_{Br} the (exogenous) reciprocity parameter of player B, K_B how kind $B's$ move is to player A, and $\hat{\lambda}_B$ how kind B believes A's first move was to player B. Denote by q the probability that player B chooses to RETURN after observing player A chose to SEND. Then, denote by α the first-order belief of player A on q, and by β the second-order belief, player B's belief on α. Following Dufwenberg and Kirchsteiger (2004), we define K_B as how much better what player B offers to player A is when compared to the average

[4] Even with an arbitrarily small probability $\epsilon > 0$ of player 1 choosing SEND, it is a best response for player 2 to also SEND.

Figure 2: Γ_3 – Psychological Trust Game with sequential reciprocity



between the highest and lowest possible efficient monetary payoff for player A, in our case $\pi^e_A = \frac{1}{2}(20 + 60) = 40$. By choosing to HOLD, player B is therefore unkind $K_B = -20$, while she is kind by returning $K_B = 20$. The perceived kindness of player A's move λ_B depends on β, the second-order belief of player B about q. If $\beta = 0$, player B believes that player A did not expect her to return, and was therefore very kind. The higher β, the less kind the move is. Given our parameter values, player A is always kind by choosing to SEND: the average between the highest and lowest B could get is $\pi^e_B = \frac{40}{2} + \frac{100-20\beta}{2} = 70 - 10\beta$, so that the perceived kindness is equal to $\hat{\lambda}_B = 100 - 20\beta - \pi^e_B = 30 - 10\beta$.

Next, we look the dynamic psychological ROSCA game. The main difference between RG and TG is that the decision to SEND in Stage 1 does not signal kindness anymore. Indeed, regardless of the probability q_i, $i \in \{1, 2\}$, that each player chooses to RETURN, both players always have a higher expected monetary payoff by choosing to SEND than to KEEP in Γ_2. Hence, the outcome in which both players KEEP and receive a payoff of 40 is inefficient in the sense of Dufwenberg and Kirchsteiger (2004). The decision to SEND being the only efficient one, the expected kindness of the first mover is always equal to zero and the Psychological ROSCA Game with Sequential Reciprocity is identical to the version with material payoffs only.[5]

Both games therefore display equilibria in which players choose to SEND with probability

[5] Using the notion of reciprocity introduced by Falk and Fischbacher (2006) (originally defined in a simultaneous setting), where kindness is defined in relative terms, we find a similar result that the decision to SEND in the first stage signals kindness in the TG, but not in the RG.

one, but the only game in which players are expected to RETURN with strictly positive probability is the TG (see Appendix A1 for a characterization of the equilibria in Γ_3).

Prediction 1. With reciprocity preferences more players choose to SEND in RG than in TG (as with self-interest), and more subjects choose to RETURN in TG.

2.3 Guilt aversion

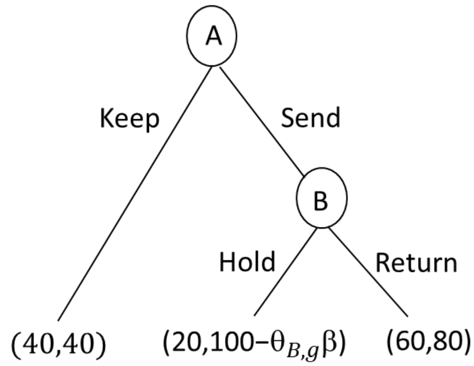
Another influential belief-based preference is that of guilt aversion (Battigalli and Dufwenberg, 2009, 2022): players bear a psychological cost from failing to meet the expectations of the others. Figure 3 presents Γ_4, the dynamic psychological TG introducing guilt aversion to the monetary payoffs shown in Γ_1. Player B has a guilt aversion parameter $\theta_{B,g}$ and second-order beliefs β on the probability α with which player A expects her to return. Hence, she prefers to return whenever $\theta_{B,g}\beta \geq 20$, and the condition for player A to prefer to SEND is $\alpha \geq 1/2$.

As shown by Attanasi et al. (2013, 2019a) the effect of the second-order beliefs on the behavior of player B in the TG goes in opposite directions with reciprocal and guilt-based preferences: while a higher β decreases the incentives to RETURN with reciprocity (because it means the first mover was not that kind), it actually increases these incentives with guilt aversion. Hence, looking at mixed strategy in Stage 2 extends the range of parameter values for which the TG displays a strictly positive probability of RETURN with reciprocity, while with guilt aversion such a mixed strategy can only emerge when a pure strategy in which player B always choose to RETURN also exists. We formally solve Γ_4 in Appendix A2.

Note that we cannot rule out a higher rate of RETURN in RG than TG with guilt aversion only: it could be true for instance that the guilt aversion parameter and the mutual expectations of cooperation are higher in the RG than in the TG.[6] As the only difference between the two games is the initial stage, such an assumption would however imply that these differences are driven by the presence of our first stage of joint cooperation. In the next

[6] Assume player i has two different guilt aversion parameters, $\theta_{i,g}^{TG}$ and $\theta_{i,g}^{RG}$ in the two games. Denote by β' the prior belief on β, before observing the first stage move. In the TG, for any value of β', we should expect player B to return whenever $\theta_{B,g}^{TG} \geq 20$, because the action of choosing SEND forces player B to update her beliefs (see Proposition 2 in Appendix A). In the RG, we should expect player i to return in the role of player B whenever $\theta_{B,g}^{RG} \geq \frac{20}{\beta'_{j,i}}$ (see Proposition 3 in Appendix A). Hence, more cooperation is possible in the RG than in the TG if the guilt aversion parameter is higher in the RG (necessary condition) and if the distribution of β' in the RG is such that enough subjects expect cooperation in the last stage.

Figure 3: Γ_4 –Psychological Trust Game with guilt aversion



section, we present what we believe is the simplest model – but not the only one – explaining how loyalty can lead to higher cooperation in the last stage.

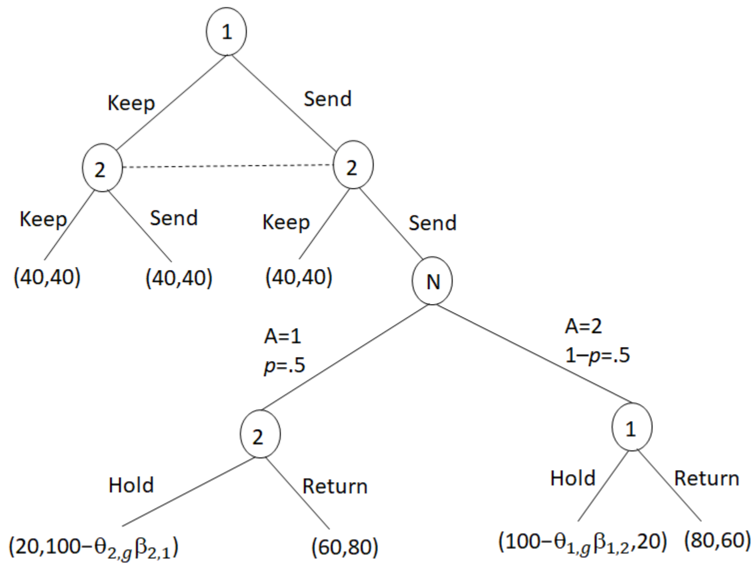
Figure 3 presents Γ_5, the RG Γ_2 as a dynamic psychological game with guilt aversion. Just as with self-interest, both players always prefer to SEND. The theoretical condition for the subject selected by Nature to take the role of player B in the last stage to prefer to RETURN is identical as in TG. The main difference however is that, when player A chooses to SEND in the first stage of the TG, the updated beliefs of player B at that node are that $\beta \geq 1/2$ (as for $\alpha < 1/2$ player A strictly prefers to KEEP). In contrast, all values of β are consistent with both players choosing SEND in the first stage of the RG. As the preference for RETURN relative to HOLD is increasing in β, for a given guilt aversion parameter and distribution of prior beliefs we would expect more subjects to RETURN in the TG than in the RG. We formally solve Γ_5 in Appendix A2.

Prediction 2. With guilt aversion of player B, more As choose to SEND in the RG, and more Bs choose to RETURN in the TG.

2.4 Loyalty

The objective of this paper is to consider the existence of a belief-dependent loyalty component in psychological utility. We seek to test the idea that individuals may develop a preference for maintaining perceived mutual cooperation, even if this perception is entirely based on beliefs. Our preferred interpretation of the concept can be expressed as an additional component to guilt aversion, based on first-order beliefs on the intention of others about their future moves

Figure 4: Psychological ROSCA Game with guilt aversion Γ_5



in the earlier stages of the game.

Perceived cooperation is defined as the belief that an action jointly taken by players had the mutual objective to obtain a Pareto-improvement over maximizing individual material payoffs only. This concept therefore includes simultaneously or sequentially taken actions leading to non-terminal nodes that allow but do not necessarily guarantee subsequent Pareto-improvements. Denote the monetary payoff of player i at a terminal node n by $\pi_{i,n}$, the set of subgame perfect Nash equilibria of the game assuming only self-interest by J, and the terminal node corresponding to one of those equilibria j by n^*_j, with $j \in J$. If there exists at least a terminal node n_k such that $\pi_{i,n_k} \geq \max_{j \in J} \pi_{i,n^*_j}$ for both $i \in \{1,2\}$ and $\pi_{i,n_k} > \max_{j \in J} \pi_{i,n^*_j}$ for at least one $i \in \{1,2\}$, denoting by K the set of nodes satisfying the condition, then a player "cooperates" to the extent that she acts under the beliefs that subsequent decisions will lead to a node $n_{k \in K}$. Every decision of the other player consistent with such beliefs can thus be perceived as cooperation.

We define *loyalty* as a preference for maintaining a history of perceived cooperation in this sense.[7] We model the preference for loyalty by assuming that betrayal incurs a psychological

[7]It is also consistent with the definition of James and Cropanzano (1994), that loyalty involves adherence to the goals of the group (in our setting, a Pareto improvement), and to the characteristic that individual

cost depending on the first-order beliefs of other players' intentions at the earlier nodes. Treason is defined as an action that is incompatible with reaching the Pareto-improving outcome. For treason to happen there must be a history of perceived cooperation in the first place.

There are in theory many different ways to define the loyalty cost of betrayal as a function of first-order beliefs on subsequent moves by co-players at a given history.[8] As our experiment consists of a two-player game with one move by the co-player in Stage 2, we can focus on this unique intention of a co-player. In the RG: whether player i, when deciding to SEND in Stage 1, was doing so with the belief player $j \neq i$ had the intention of choosing RETURN in Stage 2 if selected by Nature to take the role of player B. Denoting this belief by α_{ij}, we define the cost of betraying perceived cooperation as $\alpha_{ij}\theta_{i,l}$ where $\theta_{i,l}$ denotes player i's sensitivity to loyalty.

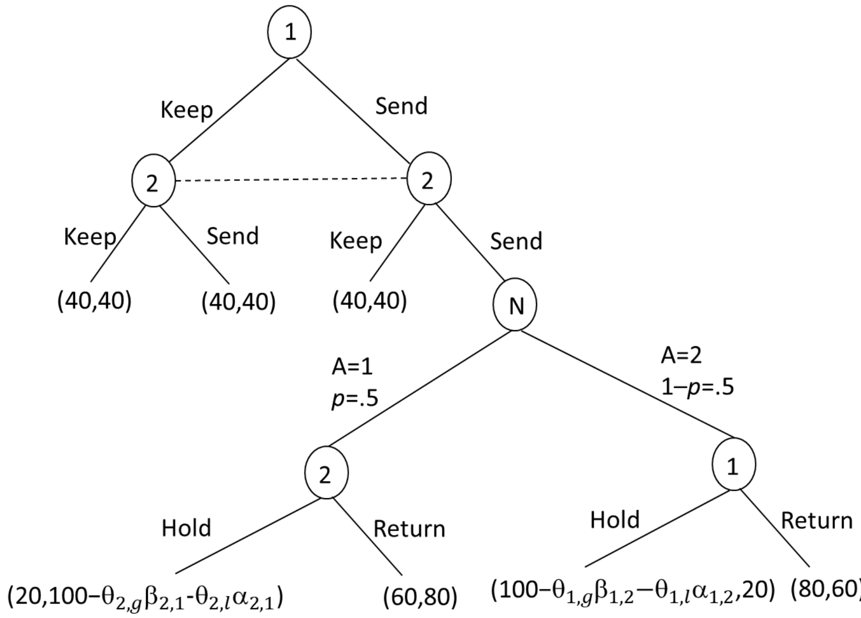
Because loyalty rests on mutual beliefs of cooperation, it is natural to combine our loyalty parameter (that depends on beliefs about the intention of the co-player) with guilt aversion (that depends on second-order belief about the intention of the player). Figure 5 presents Γ_6, a dynamic psychological ROSCA game, using the monetary payoffs shown in Γ_2. Player i has a guilt aversion parameter $\theta_{i,g}$ and second-order beliefs $\beta_{i,j}$ on the probability $\alpha_{j,i}$ with which player j expects her to return. On top of that, player i has a loyalty parameter $\theta_{i,l} \geq 0$, describing how much she dislikes choosing to HOLD if she expects player j had chosen to SEND in the first stage with the intention to RETURN if selected, measured by the first-order belief α_{ij}. Player i therefore prefers to RETURN when selected by Nature in the role of player B whenever $\theta_{i,g}\beta ij + \theta_{i,l}\alpha_{ij} \geq 20$. Hence, as compared to guilt aversion alone, the existence condition for an equilibrium in which both players choose to RETURN when selected by Nature in Stage 2 is less restrictive. Our formulation is also related to the idea of "guilt from blame" developed in Battigalli and Dufwenberg (2007), where one cares about others' inferences regarding the extent to which she is willing to HOLD in the last stage.[9]

contributions foster this sense of loyalty (Zdaniuk and Levine, 2001).

[8]Perhaps the most natural general definition is the following. Denote by S the set of players, and $s \in S$ the subset of players who move at least once and strictly more than once with strictly positive probability (necessary condition for an initial move to signal future intentions). Let $\theta_{ij,l}$ be a treason cost specific to player i's perception of joint cooperation by player j. Denote by $c_{ij}(h)$ the updated belief of player i given a history h that player $j \neq i$ had the intention to cooperate until a terminal node $n_{k\in K}$ as defined above. Player i incurs a treason cost equal to $\sum_{j\in s, j\neq i} c_{ij}(h)\theta_{ij,l}$ from making a choice at node h that does not lead to a terminal node $n_{k\in K}$ with certainty.

[9]In Battigalli and Dufwenberg (2007), "simple guilt" corresponds to the psychological cost of giving the

Figure 5: Γ_6 – Psychological ROSCA Game with guilt aversion and loyalty preferences



In contrast, the TG with guilt aversion in Γ_4 is not modified in the presence of loyalty preferences: each player moves only once, and therefore no strategy choice signals future intentions. Hence, the condition for player B to return in the dynamic psychological TG with guilt aversion and loyalty preferences are the same as with guilt aversion alone: she prefers to return whenever $\theta_{B,g}\beta \geq 20$, and the condition for player A to prefer to SEND is $\alpha \geq 1/2$. We formally solve Γ_6 in Appendix B.

Prediction 3. With loyalty of player B, more As choose to SEND in the RG than in the TG, and more (less) Bs choose to RETURN in the RG than in the TG if the positive loyalty effect in the RG is sufficiently stronger (weaker) than the positive effect of updated beliefs in the TG when B's second-order beliefs of RETURN are not lower than 50%.

Therefore, observing higher RETURN rates in the RG relative to the TG allows us to draw inference on the strong evidence of loyalty, as the increase in RG RETURN rates due to

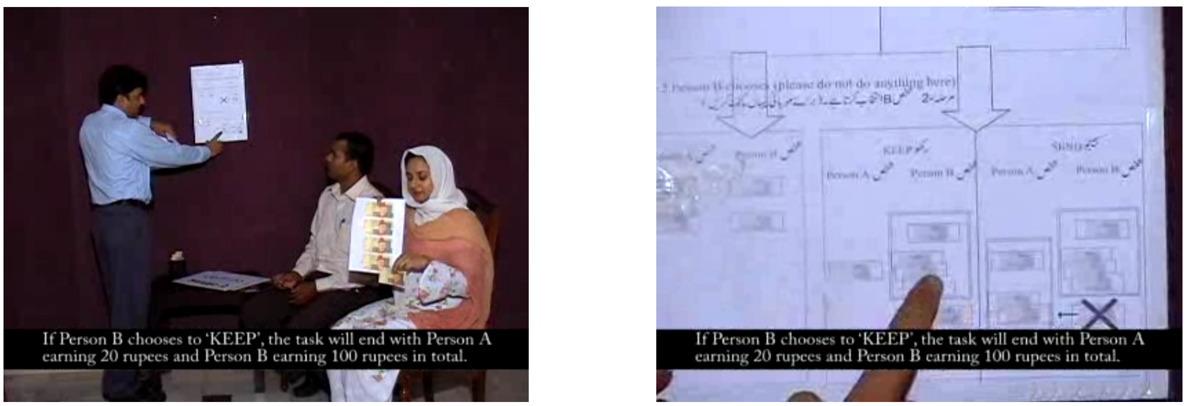
other player less than what she expected to get. In contrast, "guilt for blame" represents the psychological cost of being blamed by the other for the lower payoff they receive. In our case, it could be linked to loyalty: if you expect the other player to SEND in a spirit of future cooperation - measured by your first order belief on their probability to RETURN if playing as person B in the last stage - you also expect blame for choosing to HOLD.

loyalty not only closes the gap with TG but exceeds the increase in TG RETURN rates due to guilt aversion.

3 Experiment

3.1 Protocol

Figure 6: Experimental Instructions By Video



(a) Roles

(b) Closeup

The experiment was conducted in a textile factory in Lahore, which is the capital city of the province of Punjab, Pakistan. This is a developing country context where the ROSCA is widely used and relevant as an informal financial institution, thereby enhancing the external validity of our study. Next, as a test of loyalty as a relevant explanation of RG behavior, our lab-in-field sample departs from the non-representative (Henrich et al., 2010a) *WEIRD* Western university student sample where post-conventional moral reasoning or a training in game theory or experience in social dilemma experiments might bias a test of trust and loyalty (Henrich et al., 2010b).[10] Our subjects' ages in years range from 15 to 58 and average 31.88 ($s.d. = 9.00$), which is representative of the actual age range and mean of the working population.[11] Thus, we can test if our abstract game presented in a neutral frame and controlled

[10]For example, high socio-economic status subjects from Western countries display biases in trust games where trust and trustworthiness increase with age and stabilize around 30 (Sutter and Kocher, 2007).

[11]The legal working age in Pakistan is 15-64 years old, with a left-skewed population distribution. Based on

environment elicits loyalty compatible with the sustainability of the real-world institutions that it models. Moreover, we can enhance experimental salience and induced value by incentivizing subjects with high stakes, as reported in detail below.

Each session tested either the TG or RG, so our comparisons between these two games are between-subject. We ran four sessions per game, each with 20 subjects of two endogenously formed groups of 10 subjects each, making it 160 subjects in total for the two games.[12] The recruitment process was in itself an experimental manipulation to elicit social connections between subjects in the real world, as subjects signed up to self-selected sessions by adding their names to the lists for various sessions that were posted on a noticeboard.[13] To enhance group salience, we allowed subjects to form groups of 10 within the session seated in blocks on either side of the room through an announcement made upon arrival, "you must now allocate yourselves to two groups of 10 people per group." This was inspired by research on how endogenous choice of physical segregation reflect social distance (Schelling, 1971), e.g. cafeteria seating choices result from common group identity and norms (Ramiah et al., 2015).

This allows us to test within-subject if behavior differs when subjects play with others from within the group compared to those of a different group, similar to the endogenous formation of real world networks including ROSCA committees. Our approach is also related to previous experiments that compare interactions within and across members of natural groups such as university sports club, team, or university (Attanasi et al., 2016b); town, country, or world (Buchan et al., 2009, 2011); and groups endogenously formed in the laboratory (Arne Brekke et al., 2007; Herbst et al., 2015).[14] Leaning on previous findings of how the sense of attachment with others is more intense with social proximity (Buchan et al., 2009, 2011), the efficacy of our group manipulation is validated by comparing within-subject responses to post-experimental questions as presented below on trust attitudes towards co-participants within the group, co-participants from another group in the factory, or generally.

---

the age distribution reported by the Pakistan Bureau of Statistics (`https://www.pbs.gov.pk/sites/default/files/labour_force/publications/lfs2008_09/t01.pdf`; last accessed May 31, 2023), the estimated mean age of the urban working population is 32.1 years.

[12] One subject in TG tried to talk during the session and was asked to leave, following the rule that was announced at the beginning of the experiment.

[13] We made it clear that "You can only participate in ONE session." We checked all sign up lists to ensure that no subject has signed up for more than one session, and if those who showed up were on the list.

[14] These studies show that social proximity has positive effects on cooperation.

Subjects received task booklets, control questionnaires, and payment sheets upon being seated. A payment sheets had two halves. Each half contained a subject's uniquely assigned serial number, and either a * or ** used in conjunction with a coin flip discussed below to randomly determine their winning roles for payment purposes. Each group of 10 had five subjects with * and another five with **. Subjects entered their personally chosen passwords on both halves. They then tore off one half that was collected by the experimenter before their tasks begun, and kept the other half that was used to claim their earnings a week later; both halves had to match at the point of payment collection in order for payments to be made.

Three experimental assistants – none of whom was employed at the factory – administered each session. To guarantee uniformity and quality in delivery across sessions, subjects were instructed using pre-recorded videos of the experimental instructions (see Fig. 6). In the video recorded instructions, one person acted as the experimenter, while the other two acted as Persons A and B. The video was scripted to complement the instructions with timed gestures and reference objects. The video was filmed in spoken Urdu with English subtitles. Before subjects were allowed to make choices, we checked for their understanding with a control questionnaire. Subjects had to answer all the questions correctly before being allowed to proceed, as per standard lab protocol. Decisions were recorded on task sheets contained in their task booklets presented in both English and Urdu.[15]

Recall that sessions involved either TG or RG but not both, allowing us to make between-subject comparisons of behavior across games. Each subject was told that they had two *tasks* (i.e. games). In one task, they would be matched with someone from within the group of 10, and in the other task with someone from another group in the factory (i.e. between-group), allowing us to make within-subject comparisons of behavior across different matches. Pairwise matches were randomly determined by computer and used to calculate payments after the sessions. In half the sessions of each game, the within-group task was done before the between-group task, and counterbalanced in reverse order in the other half the sessions, with orders alternating across sessions of each game. The number of sessions, groups, and total observations per game,

[15]The video script and experimental material was originally written in English, translated to Urdu, back translated to English, and then compared to the original English version for consistency. The English versions of the experimental instructions (video script), control questionnaire, and task sheets are provided in the supplemental material. The Urdu versions are available on request. In the introductory speech, the maximum earning was erroneously stated as 300PKR rather than 100PKR. This discrepancy should not affect behavior as the game incentives were correctly stated and clearly shown in the task sheets.

role, and match are shown in table 1.

Table 1: Number of sessions, groups, and total observations per game, role, and match

	Total #subjects	Person A Group #obs	Person B Group #obs	Person A Factory #obs	Person B Factory #obs
TG (4 sessions 8 groups)	79	79	79 (53)	79	79 (43)
RG (4 sessions 8 groups)	80	80	63	80	59

Note: TG and RG were conducted between-subject. Each subject played as Person A and B in games where they were matched with co-players within and across the factory, i.e. data across roles and matches were within-subject. The number of observations as Person B by those who chose SEND as Person A in TG are enclosed in parentheses.

The games are as described in section 2, but experimentally implemented using the strategy method for comparability across TG and RG. The experimental payoffs are converted 1-to-1 as Pakistani Rupees (PKR). Each task contained two subtasks with one decision to be made in each stage (one game per role), implying a total of four decisions per subject for the two tasks. Subtasks were presented on separate decision sheets.[16] Subjects were paid on the basis of one winning task and one winning role randomly determined by two respective flips of a coin publicly observable at the end of the entire experiment. If the first coin flip lands on head (tail), then the winning task would be task 1 (2). If the second coin flip lands on head (tail), then subjects with * on their payment sheets would be paid as Person A (B) and those with ** as Person B (A). Subjects also did not receive feedback on the decisions of others or the outcome of the games during the experiment.

In the TG, subjects played both roles in each task, first as Person A and then as Person B. Person B made decisions conditional on Person A choosing SEND, but without knowing whether or not their co-player had chosen SEND or KEEP at the point of decision making. Person B's choice was considered only if Person A chose SEND. Observing SEND allows us to infer A's beliefs in the TG: if a risk-neutral A-subject sends, it is because he expects the other to return with a probability higher than half (A's first-order beliefs of RETURN). In turn, the false consensus effect (Ross et al., 1977) implies he would expect other players' second-order beliefs about his own behavior as Person B to be similar (A's third-order beliefs of RETURN).

In the RG, subjects played both stages in each task, first in Stage 1 without knowing if they would be Person A or B, then as Person B in Stage 2 without knowing if the co-player had chosen SEND or KEEP. In our analysis, a subject's choice as Person B was considered

[16] In the experiment, we kept the terms "KEEP" and "SEND" symmetric across both roles for neutral framing.

only if he had chosen SEND in the RG, for it is irrelevant if the subject had chosen KEEP in Stage 1. In calculating the payments, Person B's choice was considered only if both Persons A and B in the matched pair chose SEND. Observed actions allow us to infer loyalty in the RG: if a subject chooses SEND and RETURN if playing as person B in the last stage, it means choosing SEND was done in a spirit of future cooperation.

After subjects had completed their tasks, we collected the task booklets and administered a post-experimental questionnaire. It included questions on demographic information and trust attitudes towards people in general, the factory, and one's group of 10 and on the respondent's trustworthiness, adapted from the General Social Survey (GSS), which subjects responded to on an ordinal scale. Such responses have been shown by Glaeser et al. (2000) to be correlated with experimental trust decisions, but we mainly use them to validate the social distance between subjects of different matches in terms of broadly interpretable attitudes.

- *HELPFUL*: (Generally speaking) [As for the people in your factory] {As for the people in your group of 10}, would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?

- *TRUSTWORTHY*: (Generally speaking) [As for the people in your factory] {As for the people in your group of 10}, would you say that most people can be trusted or that you can't be too careful in dealing with people?

- *FAIR*: Do you think most people (Generally speaking) [in your factory] {in your group of 10} would try to take advantage of you if they got the chance, or would they try to be fair?

- *RECIPROCITY*: (In general)[to the people in this factory] {to the people in my group of 10}, "I am always trustworthy". Please indicate your level of agreement or disagreement with this statement.

At the end of the experiment, the payments were calculated, put into envelopes, and given to the subjects according to their serial numbers.[17] The expected payoff is 55.6PKR,

[17] To preserve anonymity and assure subjects of a fair and unbiased payment process, subjects were advised that the third party who would pay them neither worked at the factory and nor was present at the experimental sessions. This person did not know how much each subject received and was different from the person who calculated the payments.

and the maximum was 100PKR, which was worth around half a day's wage for unskilled workers (200PKR/day).

3.2 Experimental hypotheses

Given our experimental design and theoretical predictions 0-3 of sections 2.1-2.4, we can make the following experimental hypotheses. H1 and H2 directly follow from theoretical predictions 0-3.

H1: *Participation.* More subjects choose SEND in the RG than in the TG.

This hypothesis is common across all the theoretical models we have looked at: self-interest (Prediction 0), reciprocity (Prediction 1), guilt aversion (Prediction 2) and loyalty (Prediction 3).

H2: *Loyalty.* More subjects choose RETURN in the RG than in the TG.

Under reciprocity or guilt aversion, we expect a higher share of RETURN in the TG. If loyalty is sufficiently important however, we would expect more RETURN in the RG. The remaining hypotheses are based on behavioral insights (H3-H5) from the literature that are related to our experimental design. H3 and H4 follow from our meta-strategy design of the TG, i.e., all subjects play it in both roles.

H3: *Consensus effect.* In TG, subjects who SEND in the role of player A are also more likely to RETURN in the role of player B.

Subjects who RETURN expect the others act similarly and therefore SEND. In the context of guilt aversion, false consensus could lead to players having second-order beliefs that their co-players' first-order beliefs are similar to theirs (Altmann et al., 2008; Blanco et al., 2014). There is a link between this idea and our concept of loyalty: a subject in the role of player B in a game has made a privately observable first move in the role of Player A. In a sense, two subjects choosing to SEND as player A with the intent to RETURN in the role of player B also made an initial move with the intention to cooperate further. But this move is not observable, and therefore does not build a history of perceived cooperation in our definition of loyalty.

H4: *Robustness.* More subjects choose RETURN in RG than in TG, even when we restrict the comparison to those who SEND as player A in TG.

H4 would characterize the importance of loyalty if – even compared to those who expect cooperation in the TG (assuming **H4** holds) – subjects in the RG choose to RETURN more. It also helps distinguish observed cooperation in the RG from the privately observed cooperation in the TG as described in **H4**.

H5: *Social proximity.* Trust perceptions are more favorable within groups than between groups. More subjects in the TG choose SEND when playing within groups than between groups, while more subjects in the TG and RG choose RETURN when playing within groups than between groups.

This hypothesis is in line with the literature on the interplay between social proximity and trust (Glaeser et al., 2000), loyalty (Hildreth et al., 2016; Tang et al., 2017; Van Vugt and Hart, 2004), and experiments on how group biases can be driven by social preferences or guilt aversion (Charness et al., 2022; Chen and Li, 2009; Ciccarone et al., 2020).

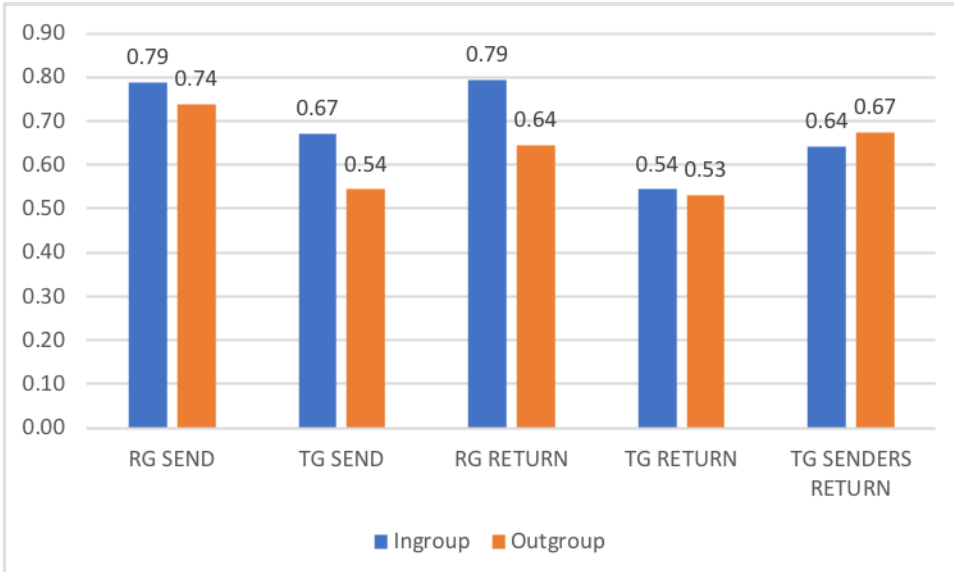
4 Results

Figure 7 shows the probability of SEND and RETURN as a function of the game played. Choosing KEEP in RG ends the game and makes Stage 2 choices redundant. Thus, we estimate the conditional probabilities of RETURN in Stage 2 of the RG only for subjects who chose SEND in Stage 1. We report TG RETURN averaged over all subjects. We also estimate the TG RETURN averages of subjects who chose SEND in Stage 1 in the role of first mover.

We observe strictly positive probabilities of SEND and RETURN in the TG, which is compatible with intention-based reciprocity and outcome-based social preferences. The positive probability of SEND in the RG is compatible with all the theories discussed, while in TG it is compatible with all but self-interest. Two-sample Wilcoxon rank-sum (Mann-Whitney) tests show that subjects SEND significantly more in RG than in TG to both the ingroup ($z = 1.65, p < 0.05$, 1-tail) and the outgroup ($z = 2.53, p < 0.01$), confirming hypothesis **H1**.

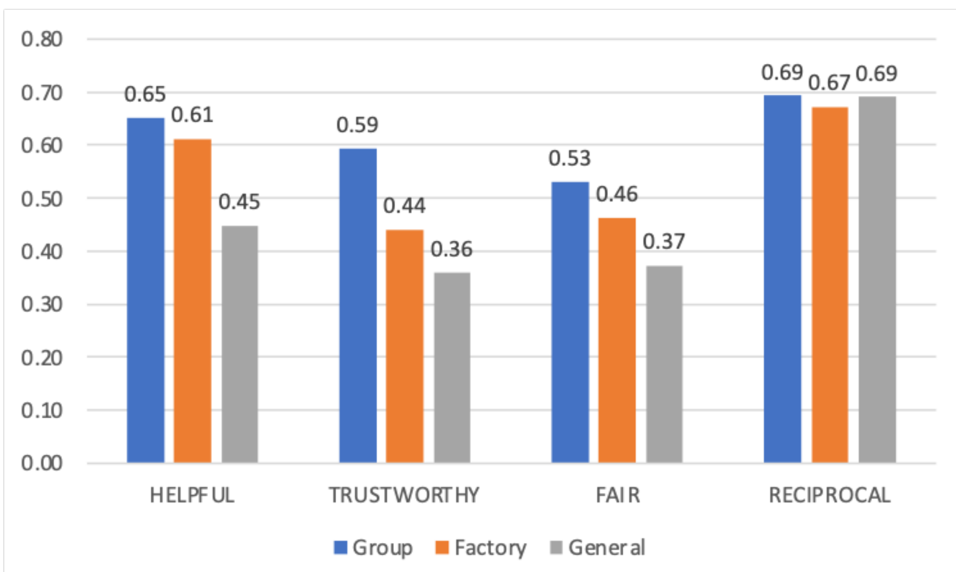
The positive probability of RETURN in the RG is compatible with guilt aversion but not with reciprocity. We observe that subjects RETURN significantly more in RG than in TG (ingroup: $z = 3.094, p = 0.001$; outgroup: $z - 1.319, p < 0.1$). This difference between treatments cannot be due to only reciprocity or guilt aversion, but is consistent with loyalty, thereby confirming **H2**.

Figure 7: Proportion of SEND and RETURN in TG and RG*



* On the horizontal axis, TG SENDERS RETURN represents the Stage 2 second mover RETURN rates of subjects who also chose SEND in the role of first movers in Stage 1 of the other game.

Figure 8: Trust Attitudes to Various Types of People*



* The scale on the vertical axis is normalized to range from 0 to 1. On the horizontal axis, labels of trust attitudes correspond to questions described in the experimental design section.

Next, we compare behavior across matching protocols, to test if loyalty is stronger when interacting within compared to across groups, which derives from the connection between group identity and loyalty as surveyed in the Introduction. First, Figure 8 validates the efficacy of our endogenous group formation manipulation in inducing trust attitudes that intensify with interactions in general, the factory, and the experimental group. We perform all within-subject comparisons using matched-pair Wilcoxon signed-rank tests. Compared to others in the factory, others within the group are thought to be more HELPFUL ($z = 2.25, p < 0.05$), TRUSTWORTHY ($z = 5.48, p = 0.0000$), FAIR ($z = 2.21, p < 0.05$), and RECIPROCAL ($z = 2.49, p < 0.001$).[18]

To test the role of second-order beliefs on guilt aversion, we match each subject's choices in both roles of the TG. In the presence of a "false consensus" effect combined with guilt aversion, one who is also willing to SEND as Person A should be loyal as Person B (one only has a move if her co-player cooperated as Person A). Therefore, we check if those who chose SEND will RETURN more frequently. When matched with someone within the group, those who chose SEND as Person A chose RETURN 64.2% of the time, while those who chose KEEP as Person A RETURN 34.5% of the time, and this difference is statistically significant ($z = 2.46, p < 0.01$). This result is robust to matching with someone from a different group, with those who SEND choosing RETURN 67.4%, versus 36.1% for those who KEEP ($z = 2.76, p < 0.01$). This validates our hypothesis **H3.**

Next, we compare our concept of loyalty with the idea of guilt driven by the false consensus effect and second-order beliefs. We find that those who chose SEND as Person A chose RETURN significantly less in the TG than in the RG ($z = -1.818, p < 0.05$) when played with someone within the group. The difference is not significant when played with someone from a different group, likely due to weaker loyalty between groups. Thus, hypothesis **H4** is confirmed only with social proximity.

Subjects SEND more to those within group (TG: $z = 1.54, p < 0.1$; RG: $z = 1.41, p < 0.1$), and RETURN more to those within group in RG ($z = 2.83, p < 0.01$) but not in TG ($n.s.$): being in the same group strengthens trust and loyalty but not reciprocity. Consistent with previous research, we find that social preferences can be enhanced by social proximity in groups

[18] These differences also hold between people in general and in the factory or group with $p < 0.001$ for all comparisons except for TRUSTWORTHY (factory: $z = 2.217, p < 0.05$) and RECIPROCAL (group: n.s.; factory: $z = -2.11, p < 0.05$).

thereby largely confirming **H5**.[19]

5 Conclusion

In this paper, we tested loyalty motives by analyzing two dynamic games for which self-interest, reciprocity, guilt aversion, and other social preference theories predict different behavioral patterns. Our main result is that a majority of subjects choose RETURN in the RG, and significantly more than in the TG. This validates loyalty as a novel preference that can complement and augment other prosocial motives. The loyalty equilibrium relies on beliefs over strategies that are never played: one adheres to cooperation as second mover when one believes the co-player also intends to adhere if assigned as second mover. More subjects SEND and RETURN to members of their endogenously formed groups, and this is consistent with the psychological literature on how loyalty prevails between socially connected individuals in groups (Van Vugt and Hart, 2004) .

Loyalty further proposes that a *shared* history of (even trivial) cooperation induces a causal effect on loyal adherence to cooperation later in the game: this our finding that more subjects who SEND *and* RETURN in the RG compared to the TG. Loyalty thus models a social mechanism through which people conditionally cooperate not only across games (Fischbacher et al., 2001) but also across stages within dynamic games (Ferguson et al., 2020; Tan et al., 2015). This stands in contrast to dynamic strategies such as *win-stay-lose-shift*, which turns exploitative by defecting once co-players cannot retaliate (Nowak and Sigmund, 1993).

Note that there are alternative specifications of loyalty that can also predict more RETURN in RG than in TG. An arguably simpler version would be to say that the RG is a different game than the TG, and that guilt aversion is more important in the RG due to loyalty concerns induced by the structure of the game itself. For example, guilt could vary across games when roles are perceived differently due to structural differences, just as how guilt varies across roles as shown by Attanasi et al., 2016a. Further studies eliciting the respective roles that beliefs and preferences play in determining the differences in behavior between the two games would help us identify the most appropriate specification of loyalty.

[19] For independent interest, we report regressions on the relationships between behavior and trust attitudes in Appendix C. We do not report them here as the layperson's interpretation of such questions is possibly too broad to pin down its role in predicting behavior as specifically defined by our theory.

Further research could improve our understanding of loyalty in at least four dimensions. First, consistent with our definition of loyalty, it has a stronger role to play when groups have a clear common goal. It would be useful to test its robustness in more complicated and repeated games. Second, we have presented the simplest possible specification for loyalty, in order to contrast it with reciprocity and guilt aversion. Loyalty can also develop from and be expressed as acts of kindness or promises. Loyalty preferences can thus be specified in conjunction with other social preferences, rather than substitute them, as has been previously suggested for others such as efficiency concerns (Engelmann and Strobel, 2004). A third avenue would be to study whether – and if so how – guilt aversion and beliefs vary across games. For example, with known guilt aversion parameters in the TG and RG, eliciting first-order and second-order beliefs would allow us to estimate the effects of guilt and loyalty on the second mover's behavior in the RG. This would require a comprehensive experimental design with multiple treatments to systematically disentangle potential confounds, considering how belief elicitation might affect decisions (Gächter and Renner, 2010) or vice versa. Fourth, looking at how much loyalty changes in different cultural contexts and frames would definitely be worthy of further investigation.

References

Altmann, S., Dohmen, T., and Wibral, M. (2008). Do the reciprocal trust less? *Economics Letters*, 99(3):454–457.

Anderson, S. and Baland, J.-M. (2002). The economics of roscas and intrahousehold resource allocation. *The Quarterly Journal of Economics*, 117(3):963–995.

Arne Brekke, K., Nyborg, K., and Rege, M. (2007). The fear of exclusion: individual effort when group formation is endogenous. *Scandinavian Journal of Economics*, 109(3):531–550.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016a). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2019a). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*, 167:341–360.

Attanasi, G., Battigalli, P., Nagel, R., et al. (2013). Disclosure of belief-dependent preferences in the trust game. *IGIER Working Paper 506, Bocconi University.*

Attanasi, G., Hopfensitz, A., Lorini, E., and Moisan, F. (2016b). Social connectedness improves co-ordination on individually costly, efficient outcomes. *European Economic Review*, 90:86–106.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2019b). Embezzlement and guilt aversion. *Journal of Economic Behavior & Organization*, 167:409–429.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature.*

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, 87:122–135.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Bracht, J. and Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39:313–326.

Brewer, M. B. and Brown, R. J. (1998). Intergroup relations. s. 554–594 in: Daniel t. gilbert, susan t. fiske und gardner lindzey (hg.): The handbook of social psychology.

Buchan, N. R., Brewer, M. B., Grimalda, G., Wilson, R. K., Fatas, E., and Foddy, M. (2011). Global social identity and global cooperation. *Psychological science*, 22(6):821–828.

Buchan, N. R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., and Foddy, M. (2009). Globalization and human cooperation. *Proceedings of the National Academy of Sciences*, 106(11):4138–4142.

Burgess, K. (1999). Loyalty dilemmas and market reform: Party-union alliances under stress in mexico, spain, and venezuela. *World Politics*, pages 105–134.

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3):560–572.

Charness, G., Di Bartolomeo, G., and Papa, S. (2022). A stranger in a strange land: Promises and identity. *Available at SSRN 4188830*.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

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

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Ciccarone, G., Di Bartolomeo, G., and Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, 124:554–568.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2019). Promises, expectations & causation. *Games and Economic Behavior*, 113:137–146.

Druckman, D. (1994). Nationalism, patriotism, and group loyalty: A social psychological perspective. *Mershon international studies review*, 38(Supplement_1):43–68.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

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

Ferguson, E., Shichman, R., and Tan, J. H. (2020). When lone wolf defectors undermine the power of the opt-out default. *Scientific Reports*, 10(1):1–12.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404.

Funk, D. C. and James, J. D. (2006). Consumer loyalty: The meaning of attachment in the development of sport team allegiance. *Journal of Sport Management*, 20(2):189–217.

Gächter, S. and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3):364–377.

Ghobadian, A., O'Regan, N., Howard, T., Gallear, D., Pesämaa, O., and Hair, J. F. (2007). More than friendship is required: An empirical test of cooperative firm strategies. *Management Decision*.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *The quarterly journal of economics*, 115(3):811–846.

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, 55(1):25–30.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010a). Most people are not weird. *Nature*, 466(7302):29–29.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Herbst, L., Konrad, K. A., and Morath, F. (2015). Endogenous group formation in experimental contests. *European Economic Review*, 74:163–189.

Hildreth, J. A. D., Gino, F., and Bazerman, M. (2016). Blind loyalty? when group loyalty makes us see evil or engage in it. *Organizational Behavior and Human Decision Processes*, 132:16–36.

Isoni, A., Sugden, R., and Zheng, J. (2022). Voluntary interaction and the principle of mutual benefit. *Journal of Political Economy*.

James, K. and Cropanzano, R. (1994). Dispositional group loyalty and individual action for the benefit of an ingroup: Experimental and correlational evidence. *Organizational Behavior and Human Decision Processes*, 60(2):179–205.

Koza, M. and Lewin, A. (2000). Managing partnerships and strategic alliances: raising the odds of success. *European Management Journal*, 18(2):146–151.

Luttmer, E. F. (2001). Group loyalty and the taste for redistribution. *Journal of political Economy*, 109(3):500–528.

Nowak, M. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432):56–58.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Ramiah, A. A., Schmid, K., Hewstone, M., and Floe, C. (2015). Why are all the w hite (a sian) kids sitting together in the cafeteria? resegregation and the role of intergroup attributions and norms. *British Journal of Social Psychology*, 54(1):100–124.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

Sherony, K. M. and Green, S. G. (2002). Coworker exchange: relationships between coworkers, leader-member exchange, and work attitudes. *Journal of applied psychology*, 87(3):542.

Stanca, L., Bruni, L., and Corazzini, L. (2009). Testing theories of reciprocity: Do motivations matter? *Journal of economic behavior & organization*, 71(2):233–245.

Sutter, M. and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic behavior*, 59(2):364–382.

Tan, J. H., Breitmoser, Y., and Bolle, F. (2015). Voluntary contributions by consent or dissent. *Games and Economic Behavior*, 92:106–121.

Tang, S., Morewedge, C. K., Larrick, R. P., and Klein, J. G. (2017). Disloyalty aversion: Greater reluctance to bet against close others than the self. *Organizational Behavior and Human Decision Processes*, 140:1–13.

Van Vugt, M. and Hart, C. M. (2004). Social identity as social glue: The origins of group loyalty. *Journal of personality and social psychology*, 86(4):585.

Zdaniuk, B. and Levine, J. M. (2001). Group loyalty: Impact of members' identification and contributions. *Journal of Experimental Social Psychology*, 37(6):502–509.

Appendix A: Resolution of the games with existing social preferences

A1 Reciprocity in the Trust Game

Proposition 1 *For intermediate values of the reciprocity parameter, $\theta_{B,r} \in (1/50, 1/40)$, there exists an equilibrium in which player A chooses SEND and player B returns with a probability $q = 3 - \frac{1}{20\theta_{B,r}} \in [1/2, 1]$. For the highest values of the reciprocity parameter $\theta_{B,r} \geq 1/40$, there exists an equilibrium in which A chooses SEND and player B chooses RETURN with probability one. For any value of the reciprocity parameter of player B, $\theta_{B,r}$, an equilibrium in which player A chooses to KEEP exists, sustained by the out-of-equilibrium belief that B would RETURN less than half of the time if player A were to SEND. This equilibrium does not survive forward induction when one of the other two equilibria exist.*

Proof.

1. The perceived kindness of player A by player B is equal to $\hat{\lambda}_B = 100 - 20\beta - \frac{40+100-20\beta}{2} = 30 - 10\beta$. The kindness of player B to player A in the second stage is either $K_B = -20$ (if she chooses to HOLD) or $K_B = 20$ if she chooses to RETURN.

2. From the perspective of player A, choosing to SEND yields higher expected payoff than choosing to KEEP if his first belief on q, $\alpha \geq \frac{1}{2}$.

3. For a given value of her second-order beliefs β, player B prefers to RETURN whenever $40\theta(30 - 10\beta) \geq 20$. Thus, in any mixed strategy with $q \leq 1$, it must hold that B is indifferent, $\alpha = \beta = q = 3 - \frac{1}{20\theta}$

4. For any value of $\theta \in [\frac{1}{40}, \frac{1}{50}]$, the above equality is satisfied with a value of $q \in [\frac{1}{2}, 1]$.

5. For values of $\theta < \frac{1}{50}$, it is satisfied with $q < \frac{1}{2}$. In that case, player A prefers to KEEP in the first stage.

6. Finally, for values of $\theta > \frac{1}{40}$, $40\theta(30 - 10\beta) \geq 20$ for all values of β, so that it is a dominant strategy for player B to RETURN regardless of β, and player A always prefers to SEND

7. For all values of θ, player A prefers to KEEP whenever $\alpha < \frac{1}{2}$. However player B must update her beliefs to $\beta \geq \frac{1}{2}$ upon observing that player A chose to SEND. When $\theta_{B,r} \geq \frac{1}{50}$, it implies that she chooses to RETURN with probability higher than $\frac{1}{2}$. Therefore, the equilibrium in which player A chooses to KEEP only survives forward induction when no other equilibrium exists.

■

A2: Guilt aversion

A2.a Trust Game

Proposition 2 *For sufficiently high guilt aversion parameter of player B, $\theta_{B,g} \geq 20$, there exists an equilibrium in which player A chooses to SEND and player B always chooses RETURN. An equilibrium in which player A chooses to KEEP always exists, sustained by the out-of-equilibrium belief that B would RETURN less than half of the time if player A were to*

SEND. It however does not survive forward induction whenever $\theta_{B,g} \geq 20$. For intermediate values of the reciprocity parameter, $\theta_{B,g} \in (20, 40)$, an equilibrium in which player A chooses to SEND and player B returns with a probability $q = \frac{20}{\theta_{B,g}} \in [1/2, 1]$ coexists with the equilibrium in pure strategy. The mixed strategy equilibrium however does not survive a small perturbation in the beliefs.

Proof.

1. If player A chooses to SEND, player B prefers to RETURN whenever $\theta_{B,g} \geq \frac{20}{\beta}$. Player A prefers to SEND whenever $20 + 40\alpha \geq 40 \Leftrightarrow \alpha > 1/2$.

2. As $\frac{20}{\beta}$ is decreasing in β, the smallest possible value of the guilt aversion parameter for B to RETURN corresponds to a second-order belief on the probability to RETURN $\beta = 1$, and is given by $\theta_{B,g} > 20$.

3. For a mixed strategy to exists, player B must be indifferent between HOLD and RETURN, $\beta = \alpha = q = \frac{20}{\theta_{B,g}}$. For this mixed strategy to be part of an equilibrium, player A must prefer to SEND. The smallest value of $\alpha = \beta = q$ such that A prefers to SEND is 1/2, so that the smallest value of $\theta_{B,g}$ such that $\frac{20}{\theta_{B,g}} \geq 1/2$ is $\theta_{B,g} \geq 40$. The highest value of $\theta_{B,g}$ such that $\frac{20}{\theta_{B,g}} < 1$ (at which point the mixed strategy degenerates to the pure strategy where B RETURNS with probability 1) is $\theta_{B,g} < 40$. This mixed strategy equilibrium is however not stable in the sense that a small tremble in those probabilities leads to best responses converging to a pure strategy equilibrium.

4. As was the case with reciprocity, the equilibrium in which player A chooses to KEEP does not survive forward induction if $\theta_{B,g} > 20$, as player A should expect B to update her beliefs to $\beta \geq 1/2$ after observing that A chose to SEND and therefore $q = \beta = 1$.

■

A2.a ROSCA Game

Proposition 3 *For any value of the guilt aversion parameters, an equilibrium in which both players chooses to SEND in the first stage exists and survives forward induction. No equilibrium with one or both players choosing to KEEP survives a small perturbation in the probabilities of the other player to SEND. When the guilt aversion parameter of a player i chosen* *by Nature to have the role of B, $\theta_{i,g} < 20$, this player chooses to HOLD in the second stage. For $\theta_{i,g} \geq 20$ two equilibrium strategies in pure strategy coexist: to RETURN or to HOLD with probability one. For $\theta_{i,g} \in (20, 40)$, an equilibrium strategy to return with a probability $q = \frac{20}{\theta_{B,g}} \in [1/2, 1]$ coexists with the equilibria in pure strategy. The mixed strategy equilibrium however does not survive a small preturbation in the beliefs.*

Proof.

- Both players prefer to SEND, even if they expect to RETURN with probability one and expect the other to HOLD with probability one.
- Each player with a guilt aversion parameter $\theta_{i,g} \geq 20$ chooses to RETURN with probability 1 when selected by Nature to have the role of B if her second-order belief about q_i is at least $\beta_{ij} \geq \frac{20}{\theta_{ij}}$.
- Hence, for all value of $\theta_{i,g} > 20$, there exists an equilibrium in which player i chooses to RETURN with probability one and one where she chooses to HOLD with probability one.
- As both players prefer to SEND regardless of q_1, q_2, the equilibrium in mixed strategy exists even for values of $q_i = \frac{20}{\theta_{ij}} < 1/2$. As for the TG, it is however not stable in the sense that a small tremble in those probabilities leads to best response converging to a pure strategy equilibrium.
- There is also no need for any player to update their beliefs upon observing the other choosing to SEND, as this strategy is compatible with any value of α. Hence, if the prior belief is that the other player will HOLD, the updated belief after observing (SEND,SEND) can remain $\beta_{ij} = 0$.

■

Comparing the TG and the RG with guilt aversion, an equilibrium in which player B (or the player chosen by Nature to take the role of player B) chooses to RETURN with probability one when $\theta_{B,g} > 20$ exists. There are two differences however. First, we should observe subjects to SEND more often in the RG. Second, we should observe subjects to RETURN more often in the TG, because after observing the other player has chosen SEND, player B updates her beliefs to $\beta \geq 1/2$, for which the only possible equilibrium value surviving a small perturbation in the beliefs is $\beta = 1$, and prefers to SEND whenever $\theta_{B,g} > 20$, while in the RG beliefs are not updated and choosing to HOLD is an equilibrium strategy even when $\theta_{B,g} > 20$.

Appendix B. Resolution of the games with Loyalty

Proposition 4 *The Trust Game with loyalty and guilt aversion displays the same equilibria as the TG with guilt aversion only.*

Proof. By definition of loyalty, a single player moves before the last stage, so that no history of perceived cooperation can be built. The game is thus identical with and without loyalty. ■

Proposition 5 *For any value of the guilt aversion parameters, an equilibrium in which both players chooses to SEND in the first stage exists and survives forward induction. No equilibrium with one or both players choosing to KEEP survives a small perturbation in the probability of the other player to SEND. There exists an equilibrium in which both players RETURN with probability one whenever $\theta_{i,g} + \theta_{i,l} \geq 20$ for both $i \in \{1, 2\}$. If this conditon does not hold, similar to the case with guilt aversion only, there exists an equilibrium in which player i RETURN with probability one regardless of what player j does whenever $\theta_{i,g} \geq 20$. The other equilibria identified in Proposition 3 continue to exist.*

Proof.

- Both players prefer to SEND, even if they expect to RETURN with probability one and expect the other to HOLD with probability one.
- Each player with a guilt aversion parameter $\theta_{i,g}$ and loyalty parameter $\theta_{i,l}$ chooses to RETURN with probability 1 when selected by Nature to have the role of B if her first-order belief about q_j, α_{ij} and her second-order belief about q_i, β_{ij} are such that $\theta_{i,g}\beta_{ij} + \theta_{i,l}\alpha_{ij} \geq 20$.
- Hence, for all value of $\theta_{i,g} + \theta_{i,l} \geq 20$ and $\theta_{j,g} + \theta_{j,l} \geq 20$, $j \neq i$, there exists an equilibrium in which $\alpha_{ij} = \beta_{ij} = q_1 = q_2 = 1$, both players choose to RETURN with probability one when selected by Nature to take the role of player B.

- When, for one of the two players, $\theta_{i,g} + \theta_{i,l} < 20$, but for the other $\theta_{j,g} \geq 20$, an equilibrium in which player i chooses to HOLD and player j to RETURN when selected by Nature exists.
- If for both players $\theta_{j,g} < 20$, only the equilibrium in which both players SEND and HOLD with probability one exists.
- There is no need for any player to update their beliefs upon observing the other choosing to SEND, as this strategy is compatible with any value of α_{ij}, β_{ij}.
- As for guilt aversion, a mixed strategy would not be stable in the sense that a small tremble in α and β leads to best response converging to a pure strategy equilibrium.

■

Comparing the Trust Game and the ROSCA game, there are two main differences. First, in the TG player B updates her beliefs on β upon observing a first move to SEND, so that $\beta \geq 1/2$, while this is not the case in the RG. For that reason, we should expect more RETURN in the TG. Second, the condition for an equilibrium in which players choose to RETURN to exist is less restrictive in the RG, because of the additional loyalty parameter. Asymmetric equilibria in which only one of the player RETURN exist under similar conditions as in the TG, but the condition for the symmetric equilibrium in which both RETURN is weaker in RG. Hence, depending on whether the first effect (updated beliefs) or the second (weaker existence condition) dominates, we should expect more or less RETURN in the TG than in the RG. As we are looking for indications that loyalty is an important feature of psychological utility, we can make the prediction that if subjects RETURN more in RG than in TG, the existence of loyalty is more important than updated beliefs.

Appendix C. Behavior and trust attitudes

Using logit regressions with standard errors clustered at the subject level, we explore the relationship between behavior in both games and the self-reported measures of perceptions regarding the trust attitudes of others. We run regressions separately for interactions within group and across the factory on SEND (= 1 if SEND, = 0 if KEEP) for player A and RETURN (= 1 if RETURN, = 0 if HOLD) for player B with the independent variables RG (= 1 if RG,

= 0 for TG), as well as trust attitudes with respects to others in the group, where HELPFUL-in, TRUSTWORTHY-in, RECIPROCITY-in, and FAIR-in (= 0 to 1), and factory, where HELPFUL-out, TRUSTWORTHY-out, RECIPROCITY-out, and FAIR-out (= 0 to 1). We control for SEX (= 1 if male, = 0 if female) and AGE (= 15-58).

Models 1 and 2 show with the positive and significant RG estimates that SEND and RETURN are higher in RG than TG when played within the group. Further, model 2 shows that RETURN increases with the perceived helpfulness of others (HELPFUL-in is positive and significant). The positive sign of RG in models 5 and 6 corroborate that subjects also SEND and RETURN more in RG than TG in outgroup interactions, and it is significant for SEND but not for RETURN, which is consistent with the weaker effect of outgroup loyalty in the RG (H2) and social proximity in the TG (H3).

Further regressions interact RG with the trust attitudes in models 3, 4, 7, and 8. We find that reciprocity increases with self-perceived trustworthiness (Model 6) but decreases with the perceived helpfulness of the outgroup (models 5 and 7) in contrast to how it increases for the ingroup (model 2). This suggests a tendency to reciprocate the ingroup but take advantage of the outgroup despite their positive intentions, which implies group-specific preference parameter values, despite no significant difference in self-reported RECIPROCITY to the ingroup and outgroup. Otherwise, the differences between effects of trust attitudes on behavior in RG and TG are not significant.

Table 2: Regressions on behavior and trust attitudes

	(1) Agp1	(2) Bgp1	(3) Agp2	(4) Bgp2	(5) Afac1	(6) Bfac1	(7) Afac2	(8) Bfac2
RG	0.797** (0.382)	1.335*** (0.407)	0.490 (1.948)	3.979 (2.698)	0.941*** (0.362)	0.588 (0.394)	1.955 (1.807)	-0.856 (1.973)
HELPFUL-in	0.225 (0.734)	1.470** (0.683)	-0.031 (0.921)	0.821 (0.835)				
TRUSTWORTHY-in	-0.441 (0.612)	-0.337 (0.607)	0.430 (0.768)	-0.320 (0.767)				
RECIPROCITY -in	-0.101 (1.288)	0.282 (1.344)	-0.947 (2.002)	1.908 (1.749)				
FAIR-in	0.486 (0.529)	-0.999* (0.521)	0.685 (0.706)	-0.436 (0.684)				
RGxHELPFUL-in			0.874 (1.739)	1.791 (1.539)				
RGxTRUSTWORTHY-in			-2.094 (1.397)	0.207 (1.279)				
RGxRECIPROCITY-in			1.860 (2.667)	-4.088 (3.324)				
RGxFAIR-in			-0.362 (1.150)	-1.667 (1.279)				
HELPFUL-out					-0.202 (0.560)	-1.623** (0.692)	-0.787 (0.927)	-2.103** (1.037)
TRUSTWORTHY-out					-0.599 (0.509)	0.080 (0.557)	-1.281* (0.776)	0.620 (0.772)
RECIPROCITY-out					-0.155 (1.275)	3.595** (1.473)	1.885 (1.969)	3.096 (1.975)
FAIR-out					-0.170 (0.504)	0.840 (0.596)	-0.438 (0.771)	0.586 (0.852)
RGxHELPFUL-out							1.131 (1.158)	1.162 (1.393)
RGxTRUSTWORTHY-out							1.395 (1.050)	-1.081 (1.104)
RGxRECIPROCITY-out							-3.919 (2.749)	1.316 (2.886)
RGxFAIR-out							0.673 (1.008)	0.698 (1.201)
SEX	0.635 (0.479)	0.697 (0.520)	0.646 (0.581)	0.950 (0.615)	0.712 (0.462)	-1.504** (0.584)	0.778* (0.469)	-1.668*** (0.615)
AGE	-0.010 (0.023)	0.022 (0.026)	-0.002 (0.022)	0.020 (0.025)	-0.024 (0.022)	0.023 (0.025)	-0.031 (0.022)	0.028 (0.026)
_cons	0.378 (1.168)	-1.558 (1.239)	0.265 (1.671)	-2.758* (1.515)	0.929 (1.176)	-1.193 (1.346)	0.484 (1.383)	-0.712 (1.564)
R^2								
N	159	142	159	142	159	138	159	138

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$