INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Disclosure of Belief-Dependent Preferences in a Trust Game

Giuseppe Attanasi, Pierpaolo Battigalli, Elena Manzoni, Rosemarie Nagel

Working Paper n. 506

This Version: August, 2022

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

Disclosure of Belief–Dependent Preferences in a Trust Game*

Giuseppe Attanasi (Sapienza University of Rome)

Pierpaolo Battigalli (Bocconi University and IGIER, Milan)

Elena Manzoni (University of Bergamo)

Rosemarie Nagel (ICREA, Universitat Pompeu Fabra, Barcelona GSE)

August 2022

Abstract

Experimental evidence suggests that agents in social dilemmas have belief-dependent, other-regarding preferences. But in experimental games such preferences cannot be common knowledge, because subjects play with anonymous co-players. We address this issue theoretically and experimentally in the context of a trust game, assuming that the trustee's choice may be affected by a combination of guilt aversion and intention-based reciprocity. We recover trustees' belief-dependent preferences from their answers to a structured questionnaire. In the main treatment, the answers are disclosed and made common knowledge within each matched pair, while in the control treatment there is no disclosure. Our main auxiliary assumption is that such disclosure approximately implements a psychological game with complete information. To organize the data, we classify subjects according to their elicited preferences, and test predictions for the two treatments using both rationalizability and equilibrium. We find that guilt aversion is the prevalent psychological motivation, and that behavior and elicited beliefs move in the direction predicted by the theory.

JEL classification: C72, C91, D03.

Keywords: Experiments, trust game, guilt, reciprocity, complete and incomplete information.

*We thank Olivier Armantier for great support in the statistical analysis. We thank for useful discussions and comments Chiara Aina, Stefania Bortolotti, Roberto Corrao, Martin Dufwenberg, Alejandro Francetich, Pierfrancesco Guarino, Andrea Guido, Sara Negrelli, Salvatore Nunnari, Fabrizio Panebianco, Jacopo Perego, Ariel Rubinstein, Fabio Tufano, and the seminar participants at University of Namur, Durham Business School, University Paris 1, Universitat Pompeu Fabra, New York University, Max Planck Institute of Economics in Jena, Sapienza University of Rome, University of Nottingham, University of Copenhagen, Bocconi University, and University of Nice. G. Attanasi gratefully acknowledges financial support by the ERC (grant DU 283953) and by "Attractivité" IDEX 2013 (University of Strasbourg). P. Battigalli gratefully acknowledges the hospitality of NYU-Stern and financial support by the ERC (grant 324219). R. Nagel gratefully acknowledges financial support by ECO2008-01768, ECO2011-25295, the Barcelona Graduate School of Economics, and the CREA program.

1 Introduction

In recent years, economists have become increasingly aware that belief-dependent motivation is important to human decision making, and that this can have important economic consequences (see, for example, Dufwenberg 2008, Battigalli & Dufwenberg 2022, and the references therein). Beliefs may affect motivation in more than one way. First, as argued by Adam Smith (1759), human action is affected by emotions and a concern for the emotions of others; since emotions can be triggered by beliefs (Elster 1998), beliefs affect choice in a non-instrumental way, that is, they affect preferences about final consequences, such as consumption allocations. Second, beliefs affect the cognitive appraisal of the pre-choice situation and the reaction to this situation, as in angry retaliations to perceived offences (Berkowitz & Harmon-Jones 2004, Battigalli *et al.* 2019b).

We study belief-dependent motivations in the Trust Game, a stylized social dilemma whereby agent A (the truster, she) takes a costly action that generates a social return, and agent B (the trustee, he) decides how to distribute the proceeds between himself and A (Berg *et al.* 1995, Buskens & Raub 2013). We focus on the simplest version of this game, called **Trust Minigame**: A can either take a costly action or not, and B can either share the proceeds equally or take everything for himself. The goal of this paper is to study, theoretically and experimentally, how B-subjects' preferences over distributions of monetary payoffs in the Trust Minigame depend on their beliefs, and how the disclosure of such belief-dependent preferences affects strategic behavior.

Two kinds of belief-dependent motivation seem salient in this social dilemma. **Guilt aversion** makes B more willing to share if he thinks that A expects him to do so; thus, B's willingness to share is *increasing* in his second-order belief, that is, B's belief that A expects B to share (Dufwenberg 2002, Battigalli & Dufwenberg 2007). On the other hand, according to **intention-based reciprocity** (see Dufwenberg & Kirchsteiger 2004), B's willingness to share depends on his perception of A's costly action as either kind or neutral toward him: The less A expects B to share, the kinder is her action; therefore, B's willingness to share is *decreasing* in his second-order belief.[1]

Experimental studies of the Trust Game find a positive correlation between elicited second-order beliefs and sharing, supporting the hypothesis that, in this social dilemma, guilt aversion is the prevailing psychological motivation of B-subjects (e.g., Charness & Dufwen-

[1]The intellectual home and mathematical framework for models of interacting agents with belief-dependent motivations is an extension of traditional game theory, put forward and labeled "psychological game theory" by Geanakoplos *et al.* (1989) and further developed by Battigalli & Dufwenberg (2009) and Battigalli *et al.* (2019a). In a nutshell, utility is assumed to depend not only on (the consequences of) choices, but also on hierarchical beliefs (see the survey by Battigalli & Dufwenberg 2022). The theory of intention-based reciprocity was first put forward by Rabin (1993) for simultaneous-move games. See also Charness & Rabin (2002), Falk & Fischbacher (2006), and Stanca *et al.* (2009).

berg 2006, Chang *et al.* 2011, and the studies surveyed in Attanasi and Nagel 2008 and Cartwright 2019).[2] Other experimental studies find evidence in support of intention-based reciprocity both in the Trust Game (Bacharach *et al.* 2007, Stanca *et al.* 2009, Toussaert 2017, Gómez-Miñambres *et al.* 2021, Rimbaud and Soldà 2021) and in other two-player social dilemmas (e.g., Falk *et al.* 2008, Dhaene & Bouckaert 2010, Dufwenberg *et al.* 2011, Dufwenberg *et al.* 2013, Chao 2018, Orhun 2018). Thus, the experimental evidence suggests that both motivations are present in social dilemmas, and especially in the role of trustee in a Trust Game.[3]

A common feature of all these experiments – and more in general of most game experiments where non-selfish preferences are likely to be important – is that such preferences are not controlled by the experimenter, hence they cannot be made common knowledge among the matched subjects. This means that the matched subjects are anonymously interacting in a game with incomplete information.[4] To see the relevance of (in)completeness of information, assume for simplicity that subjects' preferences in the Trust Game are role-dependent: A-subjects are selfish and this is common knowledge, but B-subjects are heterogeneous, as their preferences may be other-regarding in different ways and with different intensities. Suppose first that some device could make the preferences of each B-subject common knowledge within his matched pair. In such complete-information regime, information about B would work as a correlating device selecting either the cooperative outcome (when B is known to be other-regarding), or the no-trust outcome (when B is known to be selfish). In particular, we would rarely observe B grabbing the surplus created by A's costly action. Next consider the standard, incomplete-information regime: A does not know B's preferences. Since subjects are matched at random, A-subjects have to act upon beliefs about B that are necessarily independent of the true preferences of the matched B-subject. Hence, the observed joint distribution of A's and B's strategies must be (approximately) the product of the marginal distributions. Given that a fraction of A-subjects are optimistic enough to trust B, and

[2]See also Guerra & Zizzo (2004), Bacharach *et al.* (2007), Charness & Dufwenberg (2011), Bracht & Regner (2013), Ederer & Stremitzer (2017), Engler *et al.* (2018), Attanasi *et al.* (2019a, 2019b). Experimental studies of other two-player social dilemmas (Dufwenberg & Gneezy 2000, Reuben *et al.* 2009, Bellemare *et al.* 2011, Khalmetski *et al.* 2015, Khalmetski 2016, Di Bartolomeo *et al.* 2019, Attanasi *et al.* 2020, Peeters & Vorsatz 2021, Attanasi *et al.* 2022), and experimental studies of the dictator game (Balafoutas & Fornwagner 2017, Morell 2019, and Danilov *et al.* 2021) also provide support for guilt aversion.

[3]Engler *et al.* (2018) estimate the proportion of guilt and reciprocity types in a modified Trust Game where chance can stop the game after the first-mover transfer. Consistently with the findings of Attanasi *et al.* (2019b) and of this paper, the proportion of guilt types is significantly higher. However, they find a small impact on behavior of second-order beliefs, which they manipulate by changing the chance continuation probability.

[4]In a game with **complete information** there is common knowledge of (i) the rules of the game, which include how each player is paid as a function of all players' actions, and (ii) players' preferences. If at least one of these conditions fails, there is **incomplete information**. Healy (2011) finds that subjects in a laboratory experiment fail to accurately predict other subjects' preferences over possible outcomes in a set of simultaneous-move 2x2 games. Thus, such games are played in the lab with incomplete information.

a fraction of B-subjects are not trustworthy, we must observe several B-subjects grab the surplus created by A's costly action, unlike the complete-information regime.

This general result about the comparison between the predictions under complete and incomplete information can be sharpened by considering more specific assumptions about the nature of B's other-regarding preferences. If B-subjects only care about the allocation of material payoffs (e.g., because of inequity aversion, or because they maximize a weighted average of the material payoffs), then almost every B-subject must have a weakly-dominant strategy, to be carried out independently of the information regime; hence, we should observe (approximately) the same distribution of B's strategies under both complete and incomplete information. If, instead, B-subjects have belief-dependent preferences (like guilt aversion or intention-based reciprocity), then we should expect to observe different distributions under the two regimes, because the information regime should affect beliefs. But the direction and magnitude of the predicted difference depend on specific modeling choices, including the adopted solution concept.

Our study addresses these issues theoretically and experimentally: Are belief-dependent preferences heterogeneous? Are individual subjects playing the Trust Game better described by the guilt-aversion or the reciprocity model? Is it possible to disclose B's belief-dependent preferences, and do B-subjects behave as predicted given their elicited preferences? Does disclosure have the predicted impact on the behavior of matched subjects?

To answer such questions, we make the above-mentioned simplifying assumption that the truster, A, is commonly known to be self-interested; on the other hand, the trustee, B, has belief-dependent preferences given by a combination of guilt aversion and intention-based reciprocity. As in Attanasi *et al.* (2019b), we elicit the trustee's belief-dependent preferences through a structured questionnaire.[5] In the main treatment, the filled-in questionnaire is disclosed and made common knowledge within the matched pair, whereas in the control treatment, the filled-in questionnaire is not disclosed to the truster. The experimental design is such that B-subjects should not perceive an incentive to misrepresent their preferences, and indeed we find no significant difference in the pattern of answers across treatments. This supports our main auxiliary assumption: In the treatment with disclosure, B's psychological type is truthfully revealed and made common knowledge; therefore, assuming that A is commonly known to be self-interested, this treatment implements a psychological game with complete information.

[5]We explain the differences with Attanasi *et al.* (2019b) below. Bellemare *et al.* (2017, 2018) and Khalmetski *et al.* (2015) elicit the dictator's belief-dependent preferences in a dictator game through a structured questionnaire similar to ours. Regner & Harth (2014) let subjects answer to a non-structured post-experimental questionnaire (developed by psychologists) from which measures of sensitivity to guilt, positive reciprocity, and negative reciprocity are derived; they use these measures to analyze the trustee's behavior in a Trust Minigame, finding support for guilt and negative reciprocity.

To organize the data, we introduce a portable model integrating guilt aversion and reciprocity. We apply this model both to infer belief-dependent preferences from the filled-in questionnaire, and to use such "elicited" preferences to derive predictions in the Trust Minigame for the complete and incomplete-information regimes.[6] Since our subjects cannot learn from experience to play an equilibrium, we first look at the implications of rationalizability. Roughly, since A is commonly known to be selfish, in both regimes the trusting action signals a high belief that B is going to share. With this, if B is highly guilt-averse, he wants to meet A's trust; if instead B is sufficiently close to being selfish, he wants to grab all the surplus. This holds independently of the information regime. However, under complete information A knows whether one of these two cases applies, correctly predicts B's strategy, and acts accordingly. Thus, (common) knowledge of B's type allows to correlate the rationalizable actions and beliefs of the two players, whereas under incomplete information A's choice and belief are independent of B's psychological type. For intermediate types, B's strategy depends on the precise value of his second-order belief, which rationalizability does not pin down.

Hence, we refine the rationalizability predictions with equilibrium analysis, selecting the Pareto-dominating equilibrium when there are multiple equilibria. Under complete information, we obtain sharp predictions according to B's psychological type. Under incomplete information, precise Bayesian equilibrium predictions would require the specification of other parameters, such as the distribution of psychological types and interactive beliefs about such distribution (see Attanasi *et al.* 2016). To avoid arbitrary assumptions, we only provide robust qualitative predictions, which are—however—sufficient to obtain a meaningful comparison for B's behavior under the two information regimes. In particular, moderately guilt-averse types, for which rationalizability yields no prediction, tend to defect under incomplete information and to cooperate under complete information.

Experimentally, we find that guilt aversion is indeed the prevalent psychological motivation, and that behavior and elicited beliefs move in the direction predicted by the theory: First, independently of the treatment, the trustee's propensity to share is increasing with elicited guilt aversion. Second, in the treatment with disclosure there is a polarization of behavior and beliefs, with more trust and sharing in matched pairs with an elicited high-guilt trustee. Third, high-guilt trustees are less cooperative in the control (incomplete-information) treatment, where we find a higher frequency of intermediate beliefs.

As we mentioned, both in this paper and in Attanasi *et al.* (2019b) we ask B-subjects to fill in the same questionnaire, which in the main treatment is then disclosed and made public within each matched pair. Yet, in the experiment of the latter paper, subjects play a (finitely) Repeated Trust Game, while here they play a One-Shot Trust Game. Thus, we

[6] Dufwenberg *et al.* (2011) separately analyze guilt and reciprocity in a public good game.

address different questions using different theoretical approaches. For tractability reasons, in Attanasi *et al.* (2019b) we organize data by means of an equilibrium model of reputation building and neglect the reciprocity motivation. Here, instead, since we consider a simpler one-shot interaction, we can afford to jointly analyze guilt aversion and reciprocity. Furthermore, we can derive some of our behavioral predictions from 3 steps of forward-induction iterated reasoning (rationalizability), without assuming that behavior and beliefs are coordinated on an equilibrium.

The rest of the paper is structured as follows. Section 2 describes our experimental design. Section 3 presents our theoretical analysis. Section 4 presents and discusses our experimental results in light of the theoretical predictions. Section 5 concludes. An Online Appendix collects technical details about the experimental instructions (Appendix A), the theoretical analysis (Appendix B), and raw experimental data (Appendix C).

2 Design of the experiment

In this section we describe our Trust Minigame (2.1) and the experimental design (2.2), then we provide some comments (2.3). Our design is developed to test a specific model, and to estimate its parameters. We nevertheless present the design, alongside comments justifying features intuitively, before we describe the full-fledged model. We believe this provides a sense of direction before delving into the theory.

2.1 The Trust Minigame

We consider a one-shot game representing the following situation of strategic interaction (Trust Minigame): Player A ("she") and B ("he") are partners on a project with an endowment of €2. Player A has to decide whether to *Dissolve* or to *Continue* with the partnership. If player A decides to *Dissolve* the partnership, the players split the profit fifty-fifty. If player A decides to *Continue* with the partnership, total profit doubles (€4); however, in that case, player B has the right to share equally or take entirely the increased endowment. In the simultaneous-move game of Table 1 (the strategic form of the Trust Minigame), player B has to state if he would (entirely) *Take* or (equally) *Share* the higher profits before knowing

player A's choice, hence also in the case where A chooses *Dissolve*.

		B	
A		*Take*	*Share*
	Dissolve	1,1	1,1
	Continue	0,4	2,2

Table 1 Payoff matrix for the Trust Minigame.

2.2 The experimental design

Procedures Participants were first and second-year undergraduate students in Economics at Bocconi University of Milan. They voluntarily showed up at experimental sessions after having replied to E-mail or poster invitations. The sessions were conducted in a computerized classroom and subjects were seated at spaced intervals. The experiment was programmed and implemented using the z-Tree software (Fischbacher 2007). We held 16 sessions with 20 participants per session, hence 320 subjects in total. Each person could only participate in one of these sessions. Average earnings were €8.86, including a €5 show-up fee (minimum and maximum earnings were respectively €5 and €17); the average duration of a session was 50 minutes, including instructions and payment.

Design The experimental design is made of three phases and three treatments, explained in detail in Table 2 (for the experimental instructions see Appendix A). The difference between treatments concerns phase 2 and depends on whether (i) subjects playing in role B are asked to fill in a questionnaire, and (ii) such answers are disclosed within the matched A-B pair. We refer to these treatments, to be explained in detail below, as *No Q*uestionnaire (*NoQ*), *Q*uestionnaire *no D*isclosure (*QnoD*) and *Q*uestionnaire *D*isclosure (*QD*). We run 4 sessions for *NoQ* and for *QnoD* (80 subjects each) and 8 sessions for *QD* (160 subjects).

At the beginning of an experimental session, each of the 20 participants, or subjects, is randomly assigned with equal probability to role A (A-subject) or role B (B-subject) of the Trust Minigame. This determines 10 A-B pairs in each session. Each subject maintains the same role until the end of the session.

	Treatments		
	NoQ (40 pairs)	***QnoD*** (40 pairs)	***QD*** (80 pairs)
Phase 1	Trust Minigame with Beliefs Elicitation		
Phase 2	***No Q***uestionnaire	***Q***uestionnaire with ***no D***isclosure	***Q***uestionnaire with ***D***isclosure
Phase 3	Trust Minigame with Beliefs Elicitation		
	Final Questionnaire with *no* Disclosure		

Table 2 Summary of the Experimental Design.

Participants are told that the experiment is made of three phases. Instructions of each new phase are given and read aloud only prior to that phase.

We now describe in detail the three phases of the experimental design.

Phase 1 Phase 1—same for all treatments—consists of a random matching between A-subjects and B-subjects, and two subsequent decision tasks:

Belief-elicitation. With regard to the Trust Minigame of Table 1: Each A-subject is asked to guess the percentage of B-subjects in her session who will choose *Share* (A's *initial first-order belief*). Each B-subject is asked to guess the answer of his co-paired A about the percentage of B-subjects who will choose *Share* (a feature of B's *unconditional second-order belief*), and to guess the choice—*Dissolve* or *Continue*—of the co-player (a feature of B's *first-order belief*).

Choice. Within each pair, player A and player B simultaneously make their choice in the Trust Minigame of Table 1.

At the end of phase 1, subjects do not receive any information feedback about the two decision tasks. Indeed, at the beginning of this phase, they are informed that the gains in the belief-elicitation task and in the Trust Minigame will be communicated at the end of the experiment.

Phase 2 In *NoQ*, subjects proceed directly to phase 3.

In *QnoD* and *QD*, subjects are randomly re-matched to form other 10 pairs (absolute-stranger matching design). B-subjects are asked to fill in the questionnaire of Table 3 (*hypothetical payback scheme*). In particular, each B-subject is asked to consider the following hypothetical situation: His new A-co-player has chosen *Continue* and he, B, has chosen *Take*, thereby earning €4 and leaving A with €0. Given this, B has the possibility—if he wishes—to give part of this amount back to A. He is allowed to condition his payback on

the hypothesized first-order belief of A.

A thought you would have chosen *Share* with probability:	Your payback (in €):
0%	between 0.00 and 4.00
10%	between 0.00 and 4.00
20%	between 0.00 and 4.00
30%	between 0.00 and 4.00
40%	between 0.00 and 4.00
50%	between 0.00 and 4.00
60%	between 0.00 and 4.00
70%	between 0.00 and 4.00
80%	between 0.00 and 4.00
90%	between 0.00 and 4.00
100%	between 0.00 and 4.00

Table 3 Questionnaire (Hypothetical Payback Scheme) in phase 2.

Since there are 10 B-subjects, A has 11 possible guesses about how many B-subjects choose *Share* (0%, 10%, ..., 100%), which correspond to the possible beliefs shown in Table 3. Hence, each B-subject is asked to fill in each of the 11 rows of Table 3 with a value between €0.00 and €4.00. To check for framing effects, in half of the sessions of each treatment, the first column of Table 3 is shown in reverse order.

B-subjects first fill in the questionnaire on a sheet of paper and then have to copy the answers on their computer screens. A-subjects read and listen to the instructions of phase 2. Among the subjects of each *QnoD* and *QD* session, it is made public information that neither the responding B-subject nor anyone else will receive any payment for the answers to the questionnaire. Furthermore, in *QnoD* it is public information that B's filled-in questionnaire *will not be* disclosed to anyone.

On the other hand, in *QD* it is public information that B's filled-in questionnaire *will be* disclosed to a randomly-chosen A-subject. Actually, this subject is the one randomly matched with B at the beginning of phase 2. At the end of this phase, the matched B's filled-in questionnaire appears on A's screen, and the latter is invited to copy it on a sheet of paper. At this stage, subjects do not know yet that in phase 3 they are going to play again the Trust Minigame, with the same match of phase 2.

Phase 3 Phase 3—same for all treatments—consists of the same two decision tasks of phase 1, and of a new random matching. Specifically, in *NoQ* subjects are randomly re-matched to form other 10 pairs; in *QnoD* and *QD*, each A-subject is matched with the same

B-subject as in phase 2.

In *QnoD* and *QD*, each B-subject can keep his previously filled-in paper questionnaire with him for the duration of this phase. Additionally, in this phase of *QD*, A can keep the matched B's filled-in questionnaire (previously copied on a sheet of paper) with her. At the beginning of phase 3 of *QD*, it is made public information that, in each pair, B's filled-in questionnaire disclosed at the end of phase 2 corresponds to the matched B-subject of phase 3. At the end of phase 3, in *QD* and *QnoD* all filled-in questionnaires are collected by the experimenter.

Final questionnaire After phase 3, there is a final questionnaire, which is the same for all treatments (see Table 3), and equal to the one of phase 2. In *NoQ*, this is the first time B-subjects fill in the questionnaire of Table 3. In *QnoD* and *QD*, we ask B-subjects to fill in the questionnaire of Table 3 on a sheet of paper as in phase 2, knowing that it *will not be* disclosed to anyone; they may give different answers than in phase 2.

Payment Results of both phase 1 and phase 3 are communicated after the final questionnaire. In particular, each subject learns the co-player's choice in the Trust Minigame in phase 1 and in phase 3, and whether her first-order belief (A-subject) or his first and second-order beliefs (B-subject) in phase 1 and in phase 3 are correct.

2.3 Comments on the Experimental Design

In this subsection, we comment on some important features of the experimental design and provide motivations for specific design choices.

Relevance of Phase 1 There are two reasons for the initial one-shot interaction in phase 1. First, we want to know how subjects form their beliefs and make their choices without public information about B's answers to the questionnaire in phase 2. This allows us to test for within-subject effects of questionnaire disclosure in *QD*. Second, we want to let subjects understand the Trust Minigame and the belief-elicitation procedure before B-subjects fill in the questionnaire in phase 2 of *QnoD* and *QD*. Indeed, each of the 11 possible guesses for the frequency of *Share* in phase 1 corresponds to one of the 11 rows of the questionnaire of Table 3, which makes it more salient.

For *NoQ*, phase 1 has been mainly introduced to maintain the same structure as in *QnoD* and *QD*, thereby balancing observations across treatments.

Beliefs Elicitation in Phase 1 and Phase 3 We made several specific design choices about the belief-elicitation procedure, building on previous experimental literature.[7]

Charness & Dufwenberg (2006) use the strategy method to elicit the contingent choice of B-subjects in the standard, sequential version of the Trust Minigame. In this respect, our approach is similar; we make subjects play the strategic form of the Trust Minigame (see also Guerra & Zizzo 2004 and Bacharach *et al.* 2007). Due to possible framing effects, there is a subtle difference between (i) presenting subjects with a sequential game and then use the strategy method, and (ii) presenting them—as we do—with a simultaneous game corresponding to the strategic form of the sequential one (cf. Siniscalchi 2016). But we think that our description of the game in the instructions avoids such framing effects (see Appendix A).

Differently from Charness & Dufwenberg (2006), and similarly to Guerra & Zizzo (2004) and Bacharach *et al.* (2007), we elicit beliefs before choices. The experimental results by Guerra & Zizzo (2004) suggests that eliciting beliefs first should not change behavior in the subsequent Trust Minigame.

First-order beliefs of A-subjects are elicited as in Charness & Dufwenberg (2006) and follow-up papers on the Trust Minigame (see, e.g., Bracht & Regner 2013). Like them, we ask A to guess how many of the 10 B-subjects in her session would choose *Share*. Since subjects know they are paired randomly, this is a reasonable measure of first-order beliefs.

As for B-subjects, we elicit B's *unconditional* second-order belief of *Share*, while Charness & Dufwenberg (2006) elicit B's second-order belief of *Share* conditional on A choosing *Continue*. The main reason why we elicit unconditional rather than conditional beliefs relates to the questionnaire in phase 2, which has a central role in our design. As explained above, we want to match the probability grid with the 11 possible answers in A's belief-elicitation task. Thus, in order to have a manageable number of rows in the questionnaire, we only have 10 A-B pairs in each session. This is too small a number for making a reliable inference about A-subjects' first-order belief of *Share*, if one considers only those choosing *Continue*.

Our choice is also supported by theoretical considerations. On the one hand, the choice of B in the simultaneous game is equivalent to a contingent plan in the sequential version of the game, and, therefore, it should correlate with his conditional belief. On the other hand, unconditional beliefs are relevant as well, because they reflect how players reason strategically before playing the game.[8]

Our theoretical analysis of Section 3.3 provides some testable predictions about unconditional beliefs in phases 1 and 3 of *NoQ* and *QnoD* and in phase 1 of *QD*. This also motivates

[7] See Section 4 of Schotter & Trevino (2014) for a survey on first- and second-order beliefs elicitation in two-player games with belief-dependent motivations.

[8] The connection between strategic reasoning and hierarchies of initial beliefs is clarified by the literature on epistemic game theory. See the recent survey by Dekel & Siniscalchi (2015) and the references therein.

our elicitation of B's unconditional *first*-order beliefs, unlike most previous experimental studies on the Trust Minigame.[9] For the sake of simplicity, we just elicit a coarse feature of the first-order beliefs of B-subjects, that is, the action of the co-player A that they deem more likely. For the B-subjects who guess *Continue*, the unconditional second-order belief is also a rough estimate of the conditional one.[10] Notice that the payment scheme of B's second-order beliefs requires B to guess correctly both the choice and the first-order belief of A, which is consistent with the theoretical definition of second-order belief as a joint distribution about the first-order belief and the action of the co-player.

Questionnaire in Phase 2 In both treatments *QnoD* and *QD*, A-subjects read and listen to the instructions of phase 2. This is made so that A-subjects know the task of B-subjects in phase 2, and in *QD* also to help them interpret the disclosed filled-in questionnaire.

The reason for asking B-subjects to fill in the questionnaire twice—first on a sheet of paper and then on the computer screen—is to make them think more carefully about their answers. A similar consideration applies to A-subjects in *QD*: They see B's answers on their computer screen and they are asked to copy them on a sheet of paper in order to increase their attention.

Finally, we comment on withholding the identity of the recipient of B's filled-in questionnaire in the main treatment, *QD*. In phase 2, we tell subjects as little as possible about phase 3. Although subjects know that there is a phase 3, they do not know how the experiment will continue, hence they do not know if and how their answers to the questionnaire will be used later. Specifically, in phase 2 it is public information that the filled-in questionnaire *will be disclosed* to a randomly-drawn A-subject at the end of this phase. But only at the beginning of phase 3 it is made public information within each pair that the randomly-drawn player A corresponds to the matched A-subject of phase 3. With this, B-subjects should not have any obvious incentive to manipulate the beliefs of the recipient of their filled-in questionnaire.[11]

Final questionnaire When B-subjects fill in the final questionnaire, they know that there is no further decision task to execute; therefore, they should not have any incentive to lie. The final questionnaire provides information about B-subjects who did not fill in a questionnaire in phase 2 (in *NoQ*), and allows us to check whether the B-subjects who

[9] For an exception, see Regner & Harth (2014). Chang *et al.* (2011) also elicit B's first-order beliefs, although they do not use them in the analysis.

[10] Let α denote the subjective probability assigned by A to *Share*, and consider the subjective probability assigned by B to event $\alpha \leq x$, for any $x \in [0,1]$. If $\mathbb{P}_B(Cont) = 1$, then $\mathbb{P}_B(\alpha \leq x | Cont) = \mathbb{P}_B(\alpha \leq x)$.

[11] On such unexpected data use, see Charness *et al.* (2022), and, in particular, Section 5.

filled in the questionnaire in phase 2 change or confirm their answers (in *QnoD* and *QD*).[12] In the latter case, we cannot reject the hypothesis that subjects truthfully revealed their belief-dependent preferences in phase 2.

3 Model

In this section, we put forward a portable model of belief-dependent preferences with guilt aversion and intention-based reciprocity (3.1). Then we use it to derive a theoretical type-dependent payback function for the questionnaire of phase 2 (3.2), and predictions for the Trust Minigame (3.3), both under complete information (3.3.2) and incomplete information (3.3.3).

The theoretical payback function models answers to the questionnaire of phase 2; the complete-information predictions are relevant for phase 3 of *QD*; and the incomplete-information predictions are relevant for phase 1 of each treatment, and for phase 3 of *QnoD* and *NoQ* (see 3.4).

3.1 Belief-dependence, guilt, and reciprocity

We analyze the interaction of two players, i and j, who obtain monetary payoffs (m_i, m_j), and whose preferences over payoff distributions depend on beliefs. As in Battigalli & Dufwenberg (2007, 2009), we allow a player's preferences over outcomes to depend on the beliefs of the co-player, which yields a simpler representation. Higher-order beliefs appear in the expected utility-maximization problems embedded in solution concepts.

Specifically, we represent a player's preferences with a psychological utility function that depends only on (m_i, m_j) and on the co-player's first-order beliefs (which include the co-player's plan of action, a belief about what he/she is going to do). At this level of generality, we do not have to spell out the details about such beliefs. Let α_j denote j's first-order belief about the strategy pair (s_j, s_i), where the marginal on S_j represents j's plan. We obtain a utility function of the form $u_i(m_i, m_j, \alpha_j)$ by assuming that i dislikes disappointing j (the "guilt" component), and cares about the monetary payoff distribution that j expects to achieve (the "intention-based reciprocity" component); both variables depend on α_j.

We maintain the assumption that players have deterministic plans.[13] With this, let s_j be the plan (pure strategy) of player j, then α_j is determined by the pair (s_j, α_{ji}), where α_{ji} is j's

[12] In *QnoD* and *QD*, at the end of phase 3, the experimenter withdraws the phase 2 filled-in questionnaire in paper form, so as to prevent *B*-subjects from looking at their answers of phase 2 when filling in the final questionnaire. Leaving this paper with them could have biased the answers to the final questionnaire.

[13] Note that subjective expected utility maximizing players have no strict incentive to randomize. Furthermore, we find the assumption that players randomize highly problematic, unless they can actually delegate their effective action to a random device, which is not allowed in our experiment.

belief about i's strategy, and it makes sense to write $\alpha_j = (s_j, \alpha_{ji})$. For example, if A in the Trust Minigame plans to continue and expects B to share with 60% probability, then $\alpha_A = (Continue, \alpha_{AB}(Share) = 0.6)$, and her expected monetary payoff is $\mathbb{E}_A[\widetilde{m}_A; \alpha_A] = 2 \times 0.6 = 1.2$.[14] The psychological utility of B depends on this expectation. Of course, since B does not know α_A, his valuation of (m_B, m_A) is the subjective expectation $\mathbb{E}_B[u_B(m_B, m_A, \widetilde{\alpha}_A)]$ according to his second-order belief. Next we provide the details of our specification of the psychological utility function $u_B(m_B, m_A, \alpha_A)$.

The **disappointment** of player j is the difference, if positive, between j's expected payoff and his/her actual payoff: $D_j(\alpha_j, m_j) = \max\{0, \mathbb{E}_j[\widetilde{m}_j; \alpha_j] - m_j\}$.

The **kindness** of player j is the difference between the payoff that j expects to accrue to i (what j "intends" to let i have, given j's belief about i's strategy) and the "equitable" payoff of i, an average m_i^e that depends on α_{ji}: $K_j(\alpha_j) = \mathbb{E}_j[\widetilde{m}_i; \alpha_j] - m_i^e(\alpha_{ji})$.

Battigalli & Dufwenberg (2009) provide a theoretical analysis of these two belief-dependent motivations separately in Trust Minigames. We instead consider them jointly, assuming that i's preferences have an additively separable form with three terms: the utility of i's monetary payoff, the disutility of disappointing j, and the (dis)utility of increasing j's payoff if j is (un)kind. Therefore we obtain the following **psychological utility function**:

$$u_i(m_i, m_j, \alpha_j) = v_i(m_i) - g_i(D_j(\alpha_j, m_j)) + r_i(K_j(\alpha_j) \cdot m_j),\ v_i' > 0, v_i'' \leq 0, g_i' > 0, r_i' > 0. \quad (1)$$

Term $-g_i(\cdot)$ captures i's guilt aversion: i is willing to sacrifice some monetary payoff to decrease j's disappointment. Term $r_i(\cdot)$ captures i's intention-based reciprocity concerns: If j is kind (unkind), i is willing to sacrifice some monetary payoff to increase (decrease) the monetary payoff of j.

Eq. (1) is a general description of the psychological utility of a player. Next we move to the specific analysis of the Trust Minigame. We assume that preferences are role-dependent (see the discussion in Attanasi *et al.* 2016, p. 649). In particular, A (the truster) has *selfish* risk-neutral preferences, i.e., eq. (1) reduces to $u_i = m_i$. As for B, we assume that his utility may display both guilt aversion and reciprocity. We rely on belief-dependent preferences not only to analyze binary allocation choices (as, for example, in Attanasi *et al.* 2016), but also to analyze the payback scheme shown in Table 3 above, where B-subjects answer hypothetical questions by choosing distributions in a fine grid. As discussed in Attanasi *et al.* (2019b), when considered separately, guilt aversion and reciprocity have opposite effects on the payback scheme. Specifically, guilt aversion implies that the payback function is increasing in α, because the more A expects B to *Share*, the higher her expected payoff, and

[14] We use a tilde over a math symbol to denote a random variable. Because A does not know m_A after *Continue*, this number is a random variable from A's point of view, and its expectation given her first-order belief α_A is $\mathbb{E}_A[\widetilde{m}_A; \alpha_A]$.

therefore her disappointment when she receives less. Reciprocity, on the contrary, implies that the payback function is decreasing in α, as A's choice to *Continue* is expected to give more to B, hence is kinder, when α is lower.

Differently from Attanasi *et al.* (2019b), this paper analyzes, both theoretically and experimentally, the interplay between guilt aversion and reciprocity. For this reason and for the sake of tractability, in both choices that B-subjects are asked to make—in the Trust Minigame and in the hypothetical payback scheme—, we use a parametric specification of eq. (1) with the following features:

- The utility of monetary payoff, $v_i(m_i)$, is concave with constant relative risk aversion equal to 1.
- The guilt term $g_i(\cdot)$ is quadratic, as typical of most specifications of loss functions (see also Khalmetski *et al.* 2015). This assumption allows for an interior solution of the optimal payback problem.
- The reciprocity term $r_i(\cdot)$ is linear, that is, the simplest kind of odd function, as it mirrors the kindness of the other player.

To sum up, we assume the following functional form:

$$u_i(m_i, m_j, \alpha_j) = \ln(1 + m_i) - \frac{G_i}{4} \cdot [D_j(\alpha_j, m_j)]^2 + R_i \cdot K_j(\alpha_j) \cdot m_j, \tag{2}$$

where G_i and R_i respectively parametrize sensitivity to guilt and reciprocity. This parametrization achieves a good balance between tractability and flexibility.[15] Since B is the only player who may be affected by guilt and reciprocity, from now on we drop the player index from the guilt and reciprocity parameters.

In our experiment, the subjects actually play the normal form of the Trust Minigame, a simultaneous-move game (see Table 1 above). But we assume that B-subjects best respond *as if* they had observed the trusting action *Continue*, as this is the only case where their decision is relevant. This is implied by standard expected-utility maximization, except for the case where B is certain that A chooses *Dissolve*. The additional assumption is therefore that B has a belief conditional on *Continue* even when he is certain of *Dissolve*, and he acts upon such belief. Furthermore, we assume that *Continue* is regarded as fully intentional, i.e., as revealing the plan of the co-player A. The latter assumption implies that the only relevant uncertainty for B (conditional on *Continue*) is the initial belief of A about B's strategy,

[15] Jensen & Kozlovskaya (2016) provide an axiomatic analysis of guilt-averse preferences over pairs (m, g), where m is monetary payoff and g a measure of guilt. They put forward a "cancellation axiom" implying that utility is logarithmic in m, as in our model. However, their measure of guilt is (piecewise) linear, rather than quadratic.

α_{AB}. To simplify notation, from now on we let $\alpha = \mathbb{P}_A(Share)$ denote this variable, and $\beta = \mathbb{E}_B(\tilde{\alpha}|Cont)$ denote B's expectation of α, that is, the conditional second-order belief of B.

3.2 Analysis of the hypothetical payback scheme

We start with a theoretical analysis of B's answers to the questionnaire. Our baseline assumption is that B fills in the payback scheme of Table 3 as if the amount x that he hypothetically gives back to A were really given to A, thus implementing the distribution $(m_A, m_B) = (x, 4-x)$ with $x \in [0, 4]$. The expected payoff for A of action *Continue* is 2α, hence, modeling disappointment as Battigalli & Dufwenberg (2007), $D_A(\alpha, x) = \max\{0, 2\alpha - x\}$.

The kindness of action *Continue* as a function of α is modeled as in Dufwenberg & Kirchsteiger (2004), which implies that *Continue* is always a kind action, but less so the more A expects B to share (the higher α). Indeed, the higher α, the lower the increase in B's payoff that A expects to induce by choosing *Continue* rather than *Dissolve*. Specifically, the equitable payoff of B in A's eyes is the average of B's expected payoff under *Continue* and *Dissolve*: $m_B^e(\alpha) = \frac{1}{2}[\mathbb{E}_A(\widetilde{m}_B; Diss, \alpha) + \mathbb{E}_A(\widetilde{m}_B; Cont, \alpha)] = \frac{1+(4-2\alpha)}{2} = \frac{5}{2} - \alpha$; hence, the kindness of *Continue* is $K_A(\alpha) = (4 - 2\alpha) - \left(\frac{5}{2} - \alpha\right) = \frac{3}{2} - \alpha$.

Plugging $D_A(\alpha, x)$ and $K_A(\alpha)$ in (2), we obtain the maximization problem

$$\max_{x \in [0,4]} \left\{ \ln(5 - x) - \frac{G}{4} \cdot [\max\{0, 2\alpha - x\}]^2 + R \cdot \left(\frac{3}{2} - \alpha\right) \cdot x \right\}. \tag{3}$$

However, there is a possible confound. Since we put the B responder in a hypothetical situation in which he has "transgressed,"[16] we have to allow for the possibility that B chooses a higher x than implied by the solution to (3). This is because the transgression puts him in an *ex-post* negative affective state that can be alleviated by giving more than he would *ex ante*. Such "moral cleansing" (Sachdeva *et al.* 2009) is consistent with experimental findings by psychologists and economists (Ketelaar & Au 2003, Silfver 2007, and Brañas-Garza *et al.* 2013).[17] Indeed, several B-subjects in our experiment provided comments to the filled-in questionnaire in Table 2 that are in line with such repair-behavior hypothesis.[18] Therefore,

[16] Each B-subject in phase 2 is asked to consider the following hypothetical situation: "suppose that [...] A chose *Continue* and you chose *Take*, hence you got €4 and left A with €0 in his/her pocket." See the experimental instructions in *Online Appendix A*.

[17] In particular, Silfver (2007) shows that the action-tendency associated to guilt is to engage in "repair behavior." Note that, instead, the theory of guilt aversion (Dufwenberg 2002, Battigalli & Dufwenberg 2009) highlights avoidance of the anticipated negative valence associated with guilt.

[18] In *Online Appendix C* we report the answers to the debriefing questions about subjects' interpretation of their filled-in questionnaire: (a) "Explain the meaning of the values you entered in the Hypothetical Payback Scheme. Did you enter these values according to a specific feeling?" and (b) "What kind of relationship is there between this feeling and your partner's guess about you choosing Share?"

we introduce in the maximization problem an ex-*p*ost feeling-mitigation parameter $p \in [0,1]$ that boosts the payback x by adding to α in the disappointment function and subtracting from it in the kindness function. The modified maximization problem is

$$\max_{x\in[0,4]} \left\{ \ln(5-x) - \frac{G}{4} \cdot \left[\max\left\{0, 2(\alpha+p) - x\right\}\right]^2 + R \cdot \left(\frac{3}{2} - (\alpha - p)\right) \cdot x \right\}. \tag{4}$$

By strict concavity, (4) has a unique solution $x^* = \xi(\alpha)$. We call $\xi(\alpha)$ the **payback function.**[19]

Next we describe the main features of the payback function $\xi(\alpha)$ and its dependence on guilt, reciprocity, and ex-post feeling-mitigation components. Proposition 1 shows how the slope of the payback function $\xi(\alpha)$ depends on the comparison between guilt and reciprocity components. In each case, $\xi(\alpha)$ is quasi-convex, that is, either monotone or U-shaped.

Proposition 1 *Consider the range of α where an interior solution obtains (i.e., $G(p+\alpha) + R(3/2+p-\alpha) > 1/5$, $R(3/2+p-\alpha) < 1$). The payback function $\xi(\alpha)$ is*
(i) increasing *if $G > R$ and $R \leq \underline{R}(p)$,*
*(ii) first decreasing and then increasing (*U-shaped*) if $G > R$ and $\underline{R}(p) < R < \overline{R}(p)$,*
(iii) decreasing *if either $G < R$ or $R \geq \overline{R}(p)$,*
(iv) constant *if $G = R$ and $R \leq \underline{R}(p)$,*
where $\underline{R}(p) = 1/[(5-2p)(3/2+p)]$ and $\overline{R}(p) = 1/[(3-2p)(1/2+p)]$. Furthermore, $\xi(\alpha)$ is increasing in a neighborhood of α only if $\xi(\alpha) < 2p + 2\alpha$.

Proposition 1 describes the four possible shapes of the payback function. These results can be understood by interpreting the first-order condition for an interior solution in terms of the "marginal cost" and "marginal benefit" of the payback x:

$$MC(x) \equiv \frac{1}{5-x} = \frac{G}{2} \cdot \max\{0, 2p + 2\alpha - x\} + R \cdot \left(\frac{3}{2} + p - \alpha\right) \equiv MB(x). \tag{5}$$

Drawing the MC and MB schedules under different cases and tracing how their intersection is affected by parameter shifts shows how the optimal payback changes as a function of the first-order belief α and of parameter shifts.[20]

Roughly, if R is small, guilt aversion prevails on reciprocity (increasing ξ) if $G > R$, while the two psychological components balance each other if $G = R$ (constant ξ). If R has intermediate values, guilt aversion prevails on reciprocity only for high α (U-shaped ξ). In

[19] The *Online Appendix B.1* contains a derivation of the payback function $\xi(\alpha)$ in closed form; here we provide intuition.

[20] See Figure B.1 in *Online Appendix B.1.*

the remaining cases—that is, if R is large or at least $R > G$—reciprocity prevails (decreasing ξ).[21]

More formally, Proposition 1 implies that ξ is locally increasing at α (hence it is an interior solution) if and only if $G > R$ and $0 < \xi(\alpha) < 2p + 2\alpha$, which follows from the implicit function theorem: An interior solution $x^* = \xi(\alpha) \in (0, 4)$ to (4) satisfies the first-order condition (5); differentiating it, we get[22]

$$\xi'(\alpha) = \begin{cases} -R(5 - \xi(\alpha))^2 & \text{if} \quad \xi(\alpha) \geq 2p + 2\alpha, \\ \frac{2(5-\xi(\alpha))^2}{G(5-\xi(\alpha))^2+2}(G - R) & \text{if} \quad \xi(\alpha) < 2p + 2\alpha. \end{cases}$$

3.3 Theoretical predictions for the Trust Minigame

Since we assume that B chooses as if he had observed the trusting action *Continue*—the only situation in which B's choice matters—, we apply solution concepts for the sequential Trust Minigame where A moves first and B observes A's choice, a game with perfect information.[23]

We consider two situations: the complete-information regime of common knowledge of the psychological utility function u_B in (2), which we approximate in the lab in phase 3 of the main treatment (QD), and the incomplete-information case where u_B is not common knowledge, which is the standard situation in experiments. We also assume that A is commonly known to be selfish and risk neutral, which seems reasonable in the context of the Trust Minigame (see the experimental results in Section 4.2, and the discussion at the end of Section 5). Given this, part of our analysis is common to the complete and incomplete-information environments (Section 3.3.1).

In Section 3.3.2, we first provide a rationalizability analysis of the complete-information case based on forward-induction reasoning (cf. Battigalli & Dufwenberg 2009, Section 5; Battigalli *et al.* 2020). Since rationalizability does not yield sharp predictions for all possible cases (parameters of u_B), we also provide refined predictions based on Perfect Bayesian Equilibrium (PBE) analysis. It is well-known that psychological games have multiple PBE's even in situations where standard games have a unique PBE; the Trust Minigame is a case in point (Geanakoplos *et al.* 1989, Battigalli & Dufwenberg 2007, 2009). To obtain

[21] The intuition for the U-shaped ξ is that when A has low expectations (α small), the guilt aversion component of B's psychological utility has low impact and therefore the reciprocity component prevails, making payback decreasing in α even if $G > R$; but when A has high expectations (α large), since $G > R$, guilt prevails, making payback increasing in α.

[22] We can establish a link between the parametric specification of (1) considered in this paper and the "simple-guilt" model of Battigalli & Dufwenberg (2007): When R is low and $G \to \infty$, the model with reciprocity and quadratic guilt (2) yields the same payback function as the linear model with no reciprocity and sufficiently high "simple guilt."

[23] At the risk of being pedantic, let us remind the reader that "perfect information" means that players move in sequence and observe past choices, whereas "complete information" means that the rules of the game and players' preferences are common knowledge.

sharp predictions, we focus on a simple refinement: *We select the equilibrium with higher monetary payoffs*, which is the equilibrium with trust, whenever it exists. We show that this is consistent with forward-induction reasoning as captured by rationalizability, that is, whenever the latter provides a sharp prediction, it coincides with the Pareto-superior PBE prediction. In Section 3.3.3, we turn to incomplete information; in this case the behavioral predictions of rationalizability are weaker than under complete information. We refine these predictions, to some extent, by considering Bayesian Nash equilibria.

3.3.1 Rationalizability with forward induction: the first two steps

The first two steps of our analysis are based on the following assumptions:

1. **Rationality**: each player is rational, i.e., a subjective expected utility maximizer.
2. **Strong belief in rationality (Forward Induction):** each player is certain of the rationality of the co-player as long as such rationality is not contradicted by observed behavior.

The second assumption is the basic forward-induction (FI) reasoning (see Battigalli & Siniscalchi 2002, Battigalli & Dufwenberg 2009). Since we are assuming a private-values environment in which, for each player $i \in \{A, B\}$, i's utility of outcomes only depends on i's own personal traits (and possibly on the co-player's beliefs), the analysis of players' rationality is independent of whether there is complete or incomplete information. The same is true for the analysis of strong belief in rationality by player B, because in both environments he is assumed to know A's (selfish) utility function.

For the sake of simplicity and without substantial loss of generality, we also assume that there is a commonly known upper bound $L > \ln(5/3)$ on the guilt and reciprocity parameters G and R. Thus, the commonly known set of possible parameter pairs is $[0, L]^2$.

Rationality of *A* Since we are analyzing a psychological game where the utility function of B depends on the first-order beliefs of A, we use a notion of rationalizability that gives (either partial or sharp) predictions about the strategy and first-order belief of A.[24] In particular, the set of strategy-belief pairs consistent with A's rationality (assumption 1 above) is

$$P_A^1 = \left\{(s_A, \alpha) : s_A = Cont, \alpha \geq \frac{1}{2}\right\} \cup \left\{(s_A, \alpha) : s_A = Diss, \alpha \leq \frac{1}{2}\right\},$$

where $\alpha := \mathbb{P}_A[Share]$.

[24] A second reason to give predictions about (s_A, α) is that we elicit α, which is therefore "observable".

Rationality of *B* As for B, we have to consider his **psychological type** (G, R) (the parameter vector that identifies u_B) and define the set of triples $(s_B; G, R)$ consistent with assumptions 1 and 2 above. We consider predictions about $(s_B; G, R)$ because, if A thinks strategically, she forms beliefs about how s_B is related to (G, R).[25] Plugging the disappointment and kindness functions in (2), we obtain

$$u_B(m_B, m_A, \alpha) = \ln(1 + m_B) - \frac{G}{4} \cdot [\max\{0, 2\alpha - m_A\}]^2 + R \cdot \left(\frac{3}{2} - \alpha\right) \cdot m_A, \qquad (6)$$

where, conditional on *Continue*, $(m_A, m_B) = (2, 2)$ if player B chooses *Share* and $(m_A, m_B) = (0, 4)$ if he chooses *Take*. Therefore, player B chooses *Share* if and only if $\mathbb{E}_B[u_B(2, 2, \widetilde{\alpha})|Cont] \geq \mathbb{E}_B[u_B(4, 0, \widetilde{\alpha})|Cont]$ according to eq. (6), that is,

$$\frac{G}{4} \cdot \mathbb{E}_B[(2\widetilde{\alpha})^2|Cont] + 2R \cdot \left(\frac{3}{2} - \beta\right) - \ln\left(\frac{5}{3}\right) \geq 0. \qquad (7)$$

With this, we note that *we can analyze the "willingness-to-share" of B as if he were certain of A's first-order belief* α conditional on observing *Continue*. Indeed, for each conditional second-order belief of B about α, say a probability measure μ on $[0, 1]$, one can find an equivalent point belief (a Dirac measure) $\beta_\mu \in [0, 1]$ such that $\mathbb{E}_B[u_B(2, 2, \widetilde{\alpha})|Cont] \geq \mathbb{E}_B[u_B(4, 0, \widetilde{\alpha})|Cont]$ if and only if $u_B(2, 2, \beta_\mu) \geq u_B(4, 0, \beta_\mu)$. Therefore, in the analysis of rationalizability we reason as if B had a point belief $\beta \in [0, 1]$ about α conditional on *Continue* (thus, here the meaning of symbol β is a special case of the conditional expectation $\mathbb{E}_B[(\widetilde{\alpha}|Cont])$. With this, inequality (7) becomes

$$WS(\beta; G, R) := G\beta^2 - 2R\beta + 3R - \ln\left(\frac{5}{3}\right) \geq 0. \qquad (8)$$

Our analysis depends on the shape of B's **willingness-to-share function** $WS(\beta; G, R)$ implied by psychological type (G, R).[26] Clearly, *Share* is justifiable as a best reply for B of type (G, R) if $WS(\beta; G, R) \geq 0$ for some $\beta \in [0, 1]$, that is, if $\max_{\beta\in[0,1]} WS(\beta; G, R) \geq 0$;[27] similarly, *Take* is justifiable for B of type (G, R) if $\min_{\beta\in[0,1]} WS(\beta; G, R) \leq 0$. Conversely, if $\min_{\beta\in[0,1]} WS(\beta; G, R) > 0$ then *Share* is the only justifiable choice, that is, the dominant choice for (G, R); if instead $\max_{\beta\in[0,1]} WS(\beta; G, R) < 0$ then *Take* is the dominant choice for (G, R). Rationality implies that player B of type (G, R) chooses the dominant action when it exists. This gives the step-1 prediction set P_B^1.

[25] See, e.g., how rationalizability is defined in Battigalli & Siniscalchi (2002). Furthermore, we experimentally identify (G, R). Hence we can test these joint predictions.

[26] See Figure B.2 in *Online Appendix B.2*.

[27] Since *WS* is continuous, *max* and *sup* coincide.

Forward induction First, note that A's choice *Continue* is consistent with A's (selfish) rationality, because A may subjectively believe that *Share* is more likely than *Take*.[28] Therefore the assumption that B strongly believes in A's rationality implies that B is certain that $\alpha \geq 1/2$ conditional on *Continue*; formally,

$$\mathbb{P}_B\left(P_A^1|Cont\right) = \mathbb{P}_B\left(\widetilde{\alpha} \geq \frac{1}{2}|Cont\right) = 1.$$

With this, $(Share; G, R)$ is consistent with B's rationality and strong belief in A's rationality if and only if there is some $\beta \geq 1/2$ such that $WS(\beta; G, R) \geq 0$. The analogous statement with $WS(\beta; G, R) \leq 0$ holds for triple $(Take; G, R)$.[29] Let

$$P_B^{2,S} = \left\{(s_B; G, R) : \max_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) \geq 0, s_B = Share\right\}$$

and

$$P_B^{2,T} = \left\{(s_B; G, R) : \min_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) \leq 0, s_B = Take\right\},$$

then $P_B^2 = P_B^{2,S} \cup P_B^{2,T}$.

The foregoing analysis leads to a related question: When is it the case that, for B of type (G, R) who strongly believes in A's rationality, *Share* (respectively, *Take*) is the unique best reply independently of the specific belief of B? In other words, when is a strategy of B "forward-induction (FI) dominant" for psychological type (G, R)? The answer is that *Share* (respectively, *Take)* is FI-dominant for (G, R) if and only if $WS(\beta; G, R) > 0$ (respectively $WS(\beta; G, R) < 0$) *for every* $\beta \geq 1/2$, which is equivalent to $\min_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) > 0$ (respectively, $\max_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) < 0$). Thus, we obtain the following **FI-dominance regions** in the space of psychological types (G, R):

$$\mathbb{S} := \left\{(G, R) \in [0, L]^2 : \min_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) > 0\right\},$$

$$\mathbb{T} := \left\{(G, R) \in [0, L]^2 : \max_{\beta \in \left[\frac{1}{2},1\right]} WS(\beta; G, R) < 0\right\}$$

represented in Figure 1.[30] Finally, by definition, $\{Share\} \times \mathbb{S} \subset P_B^{2,S}$ and $\{Take\} \times \mathbb{T} \subset P_B^{2,T}$.

[28] Of course, such belief may be inconsistent with strategic reasoning given A's information, because rationality is only a relationship between belief and choice.

[29] Recall that we can restrict our attention to point conditional beliefs.

[30] Details about the boundaries of each region can be found in *Online Appendix B.2.*

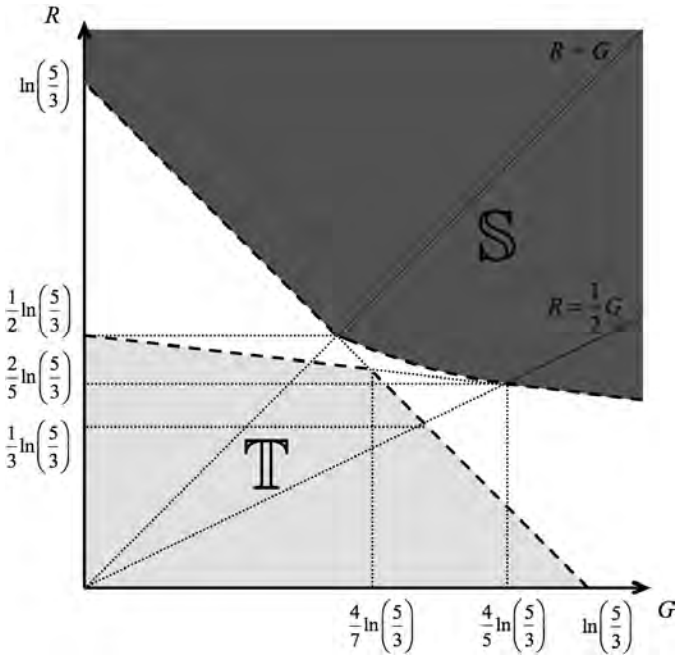

Figure 1 FI-dominance regions for *Share* and *Take* when B is certain that $\alpha \geq 1/2$.

If A assigns more than 50% probability to $\mathbb{S}$ (respectively, $\mathbb{T}$) and is certain that B satisfies assumptions 1 and 2, then $\alpha > 1/2$ (respectively, $\alpha < 1/2$).

3.3.2 Complete information

We first derive the behavioral predictions of rationalizability and then refine them by (Pareto-superior) equilibrium analysis.

Rationalizability Under complete information, the psychological type (G, R) of B is common knowledge. Therefore, rationalizability yields sharp predictions when (G, R) belongs to an FI-dominance region of Figure 1. If A believes in B's rationality and B's strong belief in rationality (assumptions 1 and 2 above), and $(G, R) \in \mathbb{S}$, then A is certain of *Share* ($\alpha = 1$) and plays *Continue*; if $(G, R) \in \mathbb{T}$, then A is certain of *Take* ($\alpha = 0$) and plays *Dissolve*. If B anticipates this and $(G, R) \in \mathbb{S}$, then he is initially certain of *Continue* and that $\alpha = 1$, and he is also certain that $\alpha = 1$ conditional on *Continue*. If instead $(G, R) \in \mathbb{T}$, then B is initially certain of *Dissolve* and that $\alpha = 0$, but strong belief in rationality implies that he would be certain that $\alpha \geq 1/2$ if he—unexpectedly—observed *Continue*.

Conversely, when the psychological type of B does not belong to any FI-dominance region (i.e., it is in the white intermediate region of Figure 1), then rationalizability does not yield predictions about strategies: even if B strongly believes in A's rationality and therefore

is certain that $\alpha \geq 1/2$ conditional on *Continue*, for each $(G, R) \notin \mathbb{S} \cup \mathbb{T}$ there is some $\beta \geq 1/2$ that makes *Share* a best reply and also some (other) $\beta \geq 1/2$ that makes *Take* a best reply. Thus, the strategy of B and the belief and strategy of A are not pinned down. The following proposition summarizes the behavioral predictions of rationalizability. Since in our experiment we do not measure the conditional second-order beliefs of B-subjects,[31] we focus on predictions about (s_A, s_B, α). When such predictions are sharp, then $\alpha \in \{0, 1\}$ and the unconditional (i.e., initial) second-order belief of B coincides with α.

Proposition 2 *Under complete information, the prediction of rationalizability based on forward induction is as follows:*
(i) Continue, Share, *and* $\alpha = 1$ *if* $(G, R) \in \mathbb{S}$,
(ii) Dissolve, Take, *and* $\alpha = 0$ *if* $(G, R) \in \mathbb{T}$,
(iii) any (s_A, s_B, α) *such that* s_A *is a best reply to* α *(i.e.,* $(s_A, \alpha) \in P_A^1$*) is possible if* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$.

Equilibrium analysis To sharpen our predictions we turn to equilibrium analysis. Since we assume that B chooses as if he had observed *Continue*, we analyze the Perfect Bayesian Equilibria (PBE) of the sequential Trust Minigame with complete information (cf. Battigalli & Dufwenberg, 2009). In a PBE initial beliefs are correct, A best responds to her initial first-order belief α, and B best responds to his conditional (second-order) belief about α, which coincides with the unconditional second-order belief when *Continue* has positive probability.[32] Specifically, (i) α, the first-order belief of A, coincides with the probability of *Share* according to the (possibly mixed) strategy of B, (ii) the unconditional belief of B assigns probability one to (s_A, α), the equilibrium strategy and first-order belief of A, (iii) if the probability of *Continue* is positive, also the conditional second-order belief of B assigns probability one to the equilibrium value α, hence $\beta = \alpha$. Mixed, or partially mixed equilibria are often justified as stable states of learning dynamics, but such justification is precluded here because we consider one-shot interactions. Pure equilibria can instead be justified (sometimes) as outcomes of strategic reasoning. Therefore, we focus on pure PBE's.

We begin with a preliminary observation. If the psychological type of B belongs to the region of the parameter space of Figure 1 where *Share* is dominant, then backward induction implies that the only PBE strategy of B is *Share*; hence, the unique PBE is the "trust equilibrium" $(Cont., Share., \alpha = \beta = 1)$. Similarly, if the psychological type of B belongs to the region where *Take* is dominant, then the only PBE strategy of B is *Take*;

[31] See the discussion in Section 2.3.

[32] Let α be the equilibrium first-order belief of A. In equilibrium, B's second-order beliefs are correct; hence, $\mathbb{P}_B[\widetilde{\alpha} = \alpha] = 1$. Since $\mathbb{P}_B[\widetilde{\alpha} = \alpha] = \mathbb{P}_B[\widetilde{\alpha} = \alpha | Cont.] \cdot \mathbb{P}_B[Cont.] + \mathbb{P}_B[\widetilde{\alpha} = \alpha | Diss.] \cdot (1 - \mathbb{P}_B[Cont.])$, if $\mathbb{P}_B[\widetilde{\alpha} = \alpha] = 1$, then either $\mathbb{P}_B[Cont.] = 0$, or $\mathbb{P}_B[\widetilde{\alpha} = \alpha | Cont.] = 1 = \mathbb{P}_B[\widetilde{\alpha} = \alpha]$.

hence, all PBE's are of the "no-trust" kind with $\alpha = 0$ and $(s_A, s_B) = (Diss, Take)$.[33] Thus, for all the aforementioned types of B, the PBE prediction is unique and coincides with the complete-information rationalizability prediction.

Now suppose that $WS(1; G, R) > 0$, but $WS(\beta; G, R) \leq 0$ for *some* $\beta < 1$. Then it is easily checked that there are multiple pure-strategy PBE's, the "trust equilibrium" and the "no-trust equilibria" mentioned above. In particular, "no-trust" is an equilibrium for each (G, R) outside the FI-dominance region $\mathbb{S}$: by definition, if $(G, R) \notin \mathbb{S}$ there is some $\beta \geq 1/2$ such that $WS(\beta; G, R) \leq 0$; hence, B is willing to *Take* even if he rationalizes the possibly unexpected choice *Continue*, which implies that there is a pure PBE of the form $(Diss, Take, \alpha = 0, \beta \geq 1/2)$ satisfying forward induction.

To obtain sharp predictions, in the case of multiplicity we apply a *Pareto-selection criterion*:[34] we assume that the pure equilibrium with higher payoffs for both players is salient and therefore players' expectations are coordinated on such equilibrium. We show that this is consistent with our complete-information rationalizability analysis based on forward-induction reasoning; hence, we are indeed refining the rationalizability predictions.

In particular, $(Cont, Share, \alpha = \beta = 1)$ is a PBE—hence the Pareto-superior equilibrium—if and only if $WS(1; G, R) \geq 0$, that is, $G + R \geq \ln(5/3) \approx 0.52$ (see eq. (8)). Therefore, if $G + R < \ln(5/3)$ ($WS(1; G, R) < 0$), "no-trust" is the unique pure PBE outcome and there is at least one such PBE that satisfies the forward-induction restriction $\beta \geq 1/2$.[35] The following proposition summarizes the Pareto-superior equilibrium predictions for s_A, s_B, and α.

Proposition 3 *The Pareto-superior, pure equilibrium prediction under complete information is as follows:*
(i) Continue, Share, *and* $\alpha = 1$ *if* $G + R \geq \ln(5/3)$,
(ii) Dissolve, Take, *and* $\alpha = 0$ *if* $G + R < \ln(5/3)$.
These predictions refine the complete-information rationalizability predictions based on forward induction.

Figure 2 builds on Figure 1; it represents the regions of the space of psychological types (G, R) with the Pareto-superior equilibrium prediction of (*Continue*,*Share*) and (*Dissolve*,*Take*) according to Proposition 3. Note that the locus of $G + R = \ln(5/3)$ is a line that

[33] The conditional second-order belief of B is arbitrary, because it is not pinned down by Bayes rule.

[34] In this psychological game, higher equilibrium material payoffs imply higher equilibrium psychological utilities.

[35] Such PBE's are not "sequential" (see Battigalli & Dufwenberg, 2009). Note also that there is a subregion below the $G + R = \ln(5/3)$ locus and above region $\mathbb{T}$ with a partially mixed PBE. In this PBE, A chooses *Continue* and B mixes with probability $\alpha^-(G, R)$, the smallest root of equation $WS(\alpha; G, R) = 0$ (in such subregion $1/2 \leq \alpha^-(G, R) < 1$).

separates the FI-dominance regions $\mathbb{S}$ and $\mathbb{T}$ in Figure 1.

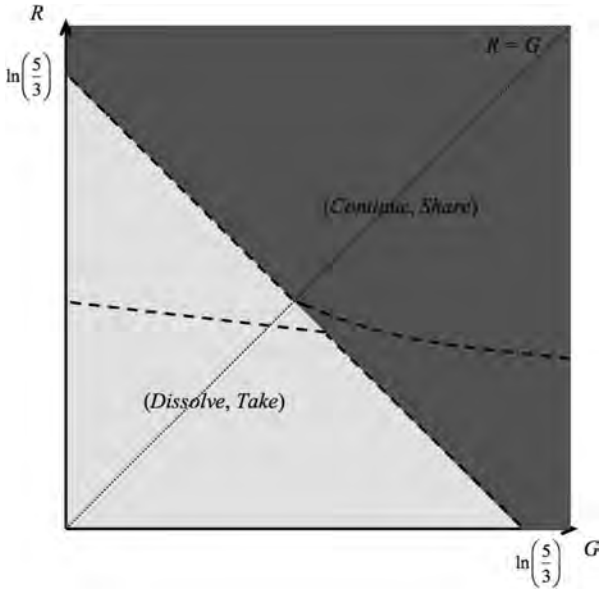

Figure 2 Pareto-superior equilibrium predictions for (s_A, s_B).

3.3.3 Incomplete information

We first derive the behavioral predictions of rationalizability, which are very coarse. As in the complete-information case and for the sake of comparison, we complement our rationalizability predictions with (Bayesian) equilibrium analysis.

Rationalizability We use a rationalizability concept for games with partially unknown utility functions, which characterizes the implications of rationality and common strong belief in rationality.[36] Steps 1 and 2 for player B are already given in Section 3.3.1: the set of possible triples $(s_B; G, R)$ consistent with rationality and strong belief in rationality is $P_B^2 = P_B^{2,S} \cup P_B^{2,T}$. Furthermore, if $(G, R) \in \mathbb{S}$ then B chooses *Share*, and if $(G, R) \in \mathbb{T}$ then B chooses *Take,* whereas if (G, R) does not belong to either FI-dominance region then both strategies can be justified by a conditional second-order belief consistent with the assumption that A is rational.

Since we are not positing any specific assumption concerning A's exogenous beliefs about the parameter vector (G, R), we cannot derive any further implication about A's behavior. To see this, note that if A assigns more than 50% probability to $\mathbb{S}$, then $\alpha > 1/2$ and the best reply is *Continue*, if instead A assigns more than 50% probability to $\mathbb{T}$, then $\alpha < 1/2$ and the best reply is *Dissolve.* Since step 3 does not refine the predictions for A, the

[36]For standard games, see Battigalli & Siniscalchi (2002) and the references therein; for psychological games, see Battigalli *et al.* (2020).

incomplete-information rationalizability algorithm stops, i.e., it gives the same predictions at each further step for each player. The following proposition summarizes:

Proposition 4 *Without restrictions on exogenous beliefs, incomplete-information rationalizability implies (only) that* $(s_A, \alpha) \in P_A^1$ *and* $(s_B; G, R) \in P_B^2$*; in particular,* B *chooses* Share *if* $(G, R) \in \mathbb{S}$ *and* Take *if* $(G, R) \in \mathbb{T}$*, while both strategies are rationalizable for* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$*.*

Equilibrium analysis We first need to introduce some terminology. We call "**exogenous**" a belief about an exogenous variable or a parameter. In particular, a belief about (G, R) is an exogenous first-order belief of A. We call "**endogenous**" a belief about a variable that we try to explain with the strategic analysis of the game. Specifically, α is the endogenous first-order belief that determines A's choice, and the cumulative functions $F_B(x) = \mathbb{P}_B(\widetilde{\alpha} \geq x)$, $F_B(\widetilde{\alpha} \geq x|Cont) = \mathbb{P}_B(\widetilde{\alpha} \geq x|Cont)$ are—respectively—the unconditional and conditional endogenous second-order beliefs of B (cf. Attanasi *et al.* 2016). Bayesian equilibrium analysis rests on specific assumptions about players' exogenous beliefs (cf. Harsanyi 1967-68). The only behavioral implications of equilibrium analysis that are robust with respect to such assumptions are those given by incomplete-information rationalizability.[37] To refine such predictions with equilibrium analysis we thus have to posit some restrictions on players' exogenous beliefs and convert them to restrictions on the distribution of behavior and endogenous beliefs.

The analysis of a fully-fledged Bayesian equilibrium model is rather complex; thus, we defer it to *Online Appendix B.2* and here we only provide a qualitative analysis based on intuition. The behavior of agents playing in role $i = A, B$ depends of their **type** t_i, which comprises their psychological type and their exogenous beliefs about the type of the co-player (exogenous higher-order beliefs). Since we assume that A is commonly known to be selfish, t_A is just a parametrization of A's exogenous hierarchy of beliefs, whereas t_B also includes the psychological parameters (G, R). With this, we describe the *equilibrium behavior and beliefs of A-types* t_A *and B-types* t_B*.*

We first list and motivate our qualitative assumptions about exogenous beliefs, anticipating some strictly related qualitative results concerning endogenous beliefs. Then we provide some intuition about key steps of the equilibrium analysis that allows us to obtain further results about behavior and endogenous beliefs. Finally, all the results are summarized in a proposition.

[37] The survey Dekel & Siniscalchi (2015) reports and explains this result for the case of games with standard preferences (see the references therein). The result can be extended to games with belief-dependent preferences.

(1) [A-heterogeneity] Since subjects cannot rely on statistical evidence on psychological types, we assume that A-subjects have *heterogeneous and dispersed exogenous first-order beliefs* about B's psychological type; specifically, the distribution across A-subjects of the expected values of G and R is dense in $[0, L]^2$. In particular, *a positive fraction of A-subjects believe that for more than half of the B-subjects it is strictly dominant to Share.* This implies that in equilibrium also endogenous first-order beliefs (the value of α for each subject) are heterogeneous and dispersed, although extreme values of α are rare, and that a positive fraction of A-subjects Continue.

(2) [B-heterogeneity] It is even more plausible that B-subjects have *heterogeneous and dispersed exogenous second-order beliefs* about the exogenous first-order beliefs of the A-subjects. Furthermore, B-subjects *believe that assumption (1) holds.* Thus, in particular, they believe that a positive fraction of A-subjects Continue. These assumptions about B-subjects imply that, in equilibrium, they have dispersed initial endogenous beliefs about the behavior and endogenous first-order belief of B, but—conditional on Continue—they are certain that $\alpha \geq 1/2$.

(3.i) [Independence between roles] When subjects are matched at random and do not observe anything about the other subject with whom they are matched, the type of A must be *independent* of the type of B. This implies that, in equilibrium, behavior and endogenous beliefs of A and B are independent as well.

(3.ii) [Independence within roles] Furthermore, we assume that the psychological type of B and his hierarchy of exogenous beliefs are independent. This implies that, in equilibrium, the psychological type of B is also independent of his endogenous beliefs.

Let $\mathbb{P}_{t_i}(\cdot)$ denote the equilibrium beliefs of type t_i of player i. Assumption (2) implies that, for each type t_B, $\mathbb{P}_{t_B}(Cont) > 0$ so that $\mathbb{P}_{t_B}(\cdot|Cont)$ is well defined. In equilibrium, an A-type t_A Continues only if $\alpha_{t_A} = \mathbb{P}_{t_A}(Share) \geq 1/2$. Therefore, for each t_B, the equilibrium conditional belief must satisfy the forward-induction requirement $\mathbb{P}_{t_B}(\widetilde{\alpha} \geq 1/2|Cont) = 1$. This in turn implies:

(4) [FI-dominance] The equilibrium predictions coincide with the rationalizability predictions of Proposition *4* for psychological types of B in the FI-dominance regions $\mathbb{S}$ and $\mathbb{T}$.

Given assumptions (1)-(2) about the dispersion of exogenous beliefs, one can also show that the distributions of α, $\mathbb{E}_B[\widetilde{\alpha}]$, and $\mathbb{E}_B[\widetilde{\alpha}|Cont]$ are dense in sub-intervals of—respectively—$[0, 1]$, $[0, 1]$, and $[1/2, 1]$, that is, there is a large fraction of subjects with "intermediate" beliefs (taking into account the forward-induction requirement for $\mathbb{E}_B[\widetilde{\alpha}|Cont]$).

The behavior of B-types t_B with psychological type (G, R) out of the FI-dominance regions depends on their equilibrium conditional belief $\mathbb{P}_{t_B}(\cdot|Cont)$.[38] Taking into account

[38]If the fraction of A-types t_A such that $\alpha_{t_A} = 1/2$ has zero measure, then $\mathbb{P}_{t_B}(\cdot|Cont) =$

that function $WS(\beta; G, R)$ is increasing on $[1/2, 1]$ if and only if $G \geq 2R$ (see eq. (8)), assumption (3.ii) implies:

(5) [Choice-belief correlation] For every psychological type (G, R) with $G \geq 2R$, a higher conditional second-order belief $F_{t_B}(\cdot|Cont)$ (in the sense of stochastic dominance) yields a higher willingness to share $\mathbb{E}_{t_B}[WS(\widetilde{\alpha}; G, R)|Cont]$.

The following proposition summarizes our qualitative predictions:

Proposition 5 *Every equilibrium of the Trust Minigame with incomplete information where a positive fraction of* A*-types choose* Continue *has the following features:*
(1) [A-heterogeneity] A*-types have heterogeneous, dispersed beliefs α about* B*'s strategy, hence, a substantial fraction of* A*-types have α well above* 0 *and well below* 1*;*
(2) [B-heterogeneity] B*-types have heterogeneous, dispersed initial beliefs about* A*'s strategy and α; conditional second-order beliefs are also heterogeneous, but have support in* $[1/2, 1]$.
(3.i) [Independence between roles] *The strategy and beliefs of* A *are independent of the strategy, psychological type, and beliefs of* B*;*
(3.ii) [Independence within roles] B*'s first- and second-order beliefs are independent of the psychological type;*
(4) [FI-dominance] B*-types with high values of* G or R *(i.e., with* $(G, R) \in \mathbb{S}$*) choose* Share*,* B*-types with low values of* G and R *(i.e., with* $(G, R) \in \mathbb{T}$*) choose* Take;
(5) [Choice-belief correlation] *The choice of intermediate types t_B depends on the equilibrium conditional belief $\mathbb{P}_{t_B}(\cdot|Cont)$; in particular, the proportion of* B*-types with $G \geq 2R$ who choose* Share *is positively correlated with the conditional second-order belief $F_{t_B}(\cdot|Cont)$.*

3.4 Theoretical predictions and experimental design

The theoretical analysis in Sections 3.3.2 (complete information) and 3.3.3 (incomplete information) leads to several testable predictions. These predictions are related to B's psychological type, elicited through the questionnaire of phase 2 (final questionnaire for NoQ). Answers to the questionnaire are supposed to reveal whether B's preferences are belief-dependent and whether guilt or reciprocity is the prevailing motivation (see Proposition 1).

Phases 1 and 3 of each treatment are meant to manipulate information about B's elicited psychological type across matched pairs as follows:

- *Phase 3 of Treatment QD:* The questionnaire filled in by B is disclosed to the matched A-subject and made common knowledge within the matched pair. Assuming that the filled-in questionnaire identifies B's psychological type and that A is commonly

$\mathbb{P}_{t_B}(\cdot|\{t_A : \alpha_{t_A} \geq 1/2\})$.

known to be selfish, the matched subjects play a psychological game with *complete information*.

- *Treatments NoQ, QnoD; Phase 1 of Treatment QD*: A obtains no information about B. Therefore the matched subjects play a psychological game with *incomplete information*.

Our testable predictions about subjects' behavior and beliefs in the Trust Minigame under the different phase-treatment combinations fall into three categories.

1. **Complete information** (phase 3 of QD): Under disclosure of the filled-in questionnaire, we predict a polarization of behavior and beliefs because common knowledge of B's psychological type works as a coordination device. If B is sufficiently selfish (low guilt and/or reciprocity parameters, $(G, R) \in \mathbb{T}$), the unique rationalizable prediction is $(Dissolve, Take)$ and $\alpha = 0$. If B is sufficiently other-regarding (high guilt and/or reciprocity parameters, $(G, R) \in \mathbb{S}$), the unique rationalizable prediction is $(Continue, Share)$, and $\alpha = 1$ (see Figure 1). Such predictions are refined by the Pareto-superior equilibrium (see Figure 2), according to which low (respectively, high) trust prevails if $G + R < \ln(5/3)$ (respectively, $G + R > \ln(5/3)$). See Propositions 2 and 3.

2. **Incomplete information** (all other phase-treatment combinations): Without disclosure, there are more heterogeneity of behavior and more dispersed beliefs. A first cause of this heterogeneity is that, by random matching, under incomplete information behavior and beliefs of A-subjects are independent of behavior and beliefs of B-subjects. As a consequence, we cannot observe the polarization on either $(Dissolve, Take)$ and $\alpha = 0$, or $(Continue, Share)$ and $\alpha = 1$, that arises under complete information.

   A second cause of heterogeneity is the presence of "intermediate" beliefs. This is quite obvious for A-subjects (assuming heterogeneous, dispersed beliefs about B's psychological type). More interestingly, there is a parameter region with intermediate values of G and low values of R ($G + R > \ln(5/3)$, $(G, R) \notin \mathbb{S}$, see Figure 1) where B-subjects would cooperate and hold high second-order beliefs under the complete-information Pareto-superior equilibrium (see Figure 2), while they exhibit less cooperative behavior and intermediate second-order beliefs under the incomplete-information Bayesian equilibrium (see Propositions 4 and 5). Symmetrically, there is also a parameter region with intermediate values of R and low values of G ($G + R < \ln(5/3)$, $(G, R) \notin \mathbb{T}$, see Figure 1) where the opposite happens, i.e., these types may cooperate under incomplete information, but not under the complete-information Pareto-superior equilibrium. We

say that **"guilt prevails for FI-underdetermined subjects"** if the fraction of subjects with utility type in the latter region is small compared to the fraction of subjects with utility type in the former region.

3. **Complete vs. incomplete information**: Rationalizability yields the same behavioral predictions (or lack thereof) for B-subjects under both complete and incomplete information (compare Proposition 2 to Proposition 4). Yet, we also consider equilibrium predictions concerning action pairs, which differ across information scenarios (compare Proposition 3 to Proposition 5). Therefore, we rely on such predictions to qualitatively compare players' behavior and beliefs across treatments and phases.

   The comparison between complete- and incomplete-information scenarios can be made *between subjects*, by comparing phase 3 of *QD vs. NoQ* and *QnoD*, and also *within subjects*, by comparing phase 3 *vs.* phase 1 of *QD*. Note that we expect no difference between phase 1 and phase 3 of *NoQ* and *QnoD*, as they both yield incomplete-information.

   First, points 1 and 2 above imply that behavior and beliefs are polarized under complete information but not under incomplete information.

   A second prediction on the comparison concerns the extent of cooperation. This crucially depends on whether guilt prevails among FI-underdetermined subjects. Since there is evidence in the literature that guilt aversion is the most important psychological motivation triggered in the Trust Minigame,[39] we expect this to be the case. Therefore, we predict more cooperative behavior of B-subjects under complete-information.

In Section 4, we discuss the data guided by the theoretical predictions for the two different information regimes.

4 Data analysis

Here we present and discuss our experimental data in light of the theoretical model. Relying on the hypothetical payback function introduced in Section 3.2, we first present in Section 4.1 the categorization of B's belief-dependent preferences derived from the answers to the questionnaire of Table 3. With this in mind, we analyze A's and B's behavior (including their side bets, hence their elicited beliefs) in the Trust Minigame using the theoretical predictions of Section 3.3. In particular, in Section 4.2 we use the complete-information predictions to analyze subjects' behavior in phase 3 of the treatment with questionnaire disclosure (QD). In Section 4.3 we use the incomplete-information predictions to analyze behavior in phase 1

[39] See Bellemare *et al.* (2017), Attanasi *et al.* (2019b), Cartwright (2019).

of *QD* and in the treatments without questionnaire disclosure (*NoQ* and *QnoD*). In Section 4.4, we compare behavior in all these phase-treatment combinations with behavior in phase 3 of *QD*.

4.1 Elicitation of belief-dependent preferences through the filled-in questionnaire

The experimental elicitation of B's belief-dependent preferences in the Trust Minigame relies on his answers to the questionnaire of Table 3 (see Section 3.2). We call "**payback pattern**" the actual answers of a B-subject, with one payback value for each hypothesized α (A's belief about B's strategy *Share*). Recall that the payback pattern gives 11 observations for B's payback function, i.e., one for each $\alpha \in \{0, 10\%, ..., 100\%\}$. In *Online Appendix C* we report payback patterns of the 160 B-subjects in our experiment.

The left panel of Figure 3 shows B-subjects' average payback pattern, disentangled by treatment. As the figure suggests, there are no treatment differences: We performed a Kruskal-Wallis test of the equality of distributions of payback values in the three treatments for each one of the 11 hypothesized α's and found a smallest *P-value* $= 0.346$ for $\alpha = 0.9$. A Mann-Whitney test with a pairwise comparison between treatments confirms this result (smallest *P-value* $= 0.165$ for $\alpha = 0.9$ in *QnoD vs. QD*).

Recall that, to test the hypothesis that B-subjects truthfully revealed their belief-dependent preferences in phase 2 of *QnoD* and *QD*, they were asked to fill in again the questionnaire at the end of the experiment (cf. Table 2). With very few exceptions, B-subjects confirmed the payback pattern of phase 2.[40] Therefore, for these two treatments we only referred to the questionnaire in phase 2, while for treatment *NoQ* we relied on the final questionnaire—the only one filled in by B-subjects in this treatment. Furthermore, we checked for each treatment that there is no framing effect on the payback due to the presentation of the 11 lines of the questionnaire in reverse order in half of the experimental sessions of each treatment (Mann-Whitney test, smallest *P-value* $= 0.129$ for $\alpha = 0.9$ in *QD*). This is confirmed by a similar ratio of increasing over decreasing payback patterns in each order of presentation (χ^2 test, *P-value* $= 0.276$).

The left panel of Figure 3 shows that average payback patterns are increasing. This is the result of the prevalence of subjects whose elicited preferences display guilt aversion, as shown in the right panel of Figure 3. Specifically, with reference to Proposition 1, we find that 138/160 (86%) B-subjects have a payback pattern that mimics one of the possible quasi-

[40] Only 9/80 (4/40) B-subjects provided a different payback pattern in the final questionnaire in *QD* (*QnoD*), and only for 3/9 (1/4) would this difference change their assignment to a category of psychological types (see Table 4). We acknowledge that this tendency of B-subjects to provide the same answers in phase 2 and the final questionnaire could be due to a consistency motive (see, e.g., Podsakoff *et al.* 2003).

convex shapes of the payback function $\xi(\alpha)$; this fraction is treatment-independent (35/40 in *NoQ*, 33/40 in *QnoD*, 70/80 in *QD*).[41] Considering only the 138 B-subjects qualitatively captured by our model, the right panel of Figure 3 reports, for each possible theoretical shape of $\xi(\alpha)$, their average payback pattern and the corresponding number of B-subjects: guilt prevails for the overwhelming majority of these B-subjects (81/138).

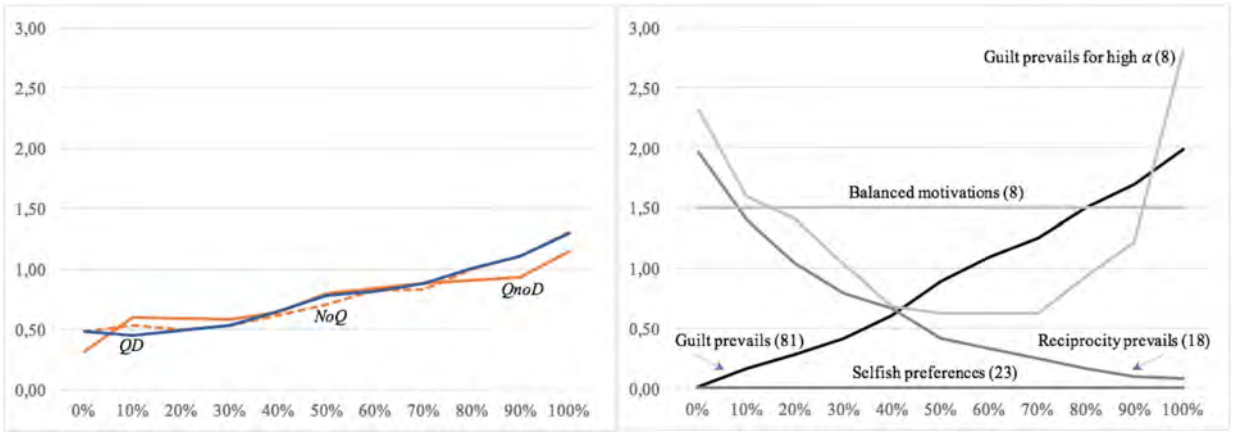

Figure 3 B's average payback pattern, by treatment (left panel) and type (right panel).
The figure reports on the left panel B's average payback pattern in *NoQ* (40 subjects), *QnoD* (40 subjects), and *QD* (80 subjects). On the right panel, it reports the average payback pattern of B-subjects according to the shapes of $\xi(\alpha)$ predicted by Proposition 1: guilt prevails ($\xi(\alpha)$ increasing), guilt prevails for high α (U-shaped $\xi(\alpha)$), reciprocity prevails ($\xi(\alpha)$ decreasing), balanced motivations ($\xi(\alpha)$ constant), and selfish preferences ($\xi(\alpha) = 0$) as a separate special case of balanced motivations; for each average pattern, the intensity of the black color indicates the relative frequency of the corresponding shape in the population of B-subjects (reported in parentheses).

Our model also allows to estimate, for each B-subject, the triple (G, R, p) that identifies B's best response to each hypothesized α, i.e., his theoretical payback function $\xi(\alpha; G, R, p)$. We denote by $\hat{G}$, $\hat{R}$, and $\hat{p}$, the estimated values of G, R, and p, respectively. The main goal of this estimation is to describe each B-subject by his estimated psychological type $(\hat{G}, \hat{R})$. This is a preliminary step to test the theoretical predictions of Propositions 2-5, which rely on the different regions of psychological types (G, R) in Figures 1-2. We also use the estimated parameters $\hat{G}$ and $\hat{R}$ to include in one of the five categories of the right panel of Figure 3 also the 22/160 (14%) B-subjects whose payback pattern does not fit any of the corresponding shapes of $\xi(\alpha)$ (for a similar method, see Costa-Gomes *et al.* 2001).

[41] In *Online Appendix C* we report B-subjects' answers to debriefing questions about the interpretation of the filled-in questionnaire.

The best-fit response function $\hat{\xi}(\alpha) := \xi(\alpha; \hat{G}, \hat{R}, \hat{p})$ of a given B-subject minimizes the sum of the squared deviations of the theoretical payback function from the payback pattern for the 11 rows of the filled-in questionnaire. Given that the maximization problem (4) is non-linear in one of the unknown parameters, G, we use non-linear least square estimation, with bounds given by $0 \leq G, R \leq 1000$ and $0 \leq p \leq 1$. To account for the small size of the sample, standard deviations are given by a (non parametric) bootstrap estimation of size 10,000. In *Online Appendix C*, we provide the non-linear least square estimates $\hat{G}$, $\hat{R}$ and $\hat{p}$ (and standard deviations associated to each estimated parameter) for the 160 B-subjects in our experiment. Across all 160 B-subjects, we find that 123 have $\hat{G} > 0$, 101 have $\hat{R} > 0$ (88 have both $\hat{G} > 0$ and $\hat{R} > 0$), and 125 have $\hat{p} > 0$, with no significant treatment difference in the distribution of each of the three estimated parameters (Kruskal-Wallis test, *P-value* $= 0.358$ for G, 0.760 for R, 0.790 for p).

In Table 4, we report the distribution of the 160 B-subjects' estimated psychological types across the five possible shapes of the corresponding payback function $\xi(\alpha)$ of Proposition 1 (categories of psychological types). The number of B-subjects whose payback pattern is not qualitatively captured by the five predicted shapes—a total of 22—is reported in parentheses.[42]

A clarification about estimated values of the ex-*p*ost feeling mitigation parameter is in order. Notice that estimated parameter $\hat{p}$ does not play any role in Table 4. The reason is that parameter p has been introduced in the maximization problem (4) to take into account the possible confound of ex-post feeling mitigation, which matters in the analysis of the hypothetical payback and in the estimation of G and R for each B-subject, but not in the analysis of the Trust Minigame. Indeed, we checked that $\hat{p}$ is independent of B-subjects' behavior in each Trust Minigame played during the experiment, and of A-subjects' behavior in the Trust Minigame of phase 3 of *QD* (i.e., under disclosure).

Table 4 shows no significant difference between the distributions of types in *NoQ* and *QnoD* (χ^2 test, *P-value* $= 0.639$), which allows us to pool the data of these two treatments (column *NoQ-QNoD* in Table 4) so as to have the same number of observations without disclosure (*NoQ-QnoD*) and with disclosure (*QD*). Table 4 also shows no significant difference between the distributions of psychological types in *NoQ-QNoD* and *QD* (last two columns of Table 4: χ^2 test, *P-value* $= 0.734$). This is further evidence that the presence or absence of information disclosure does not affect subjects' answers to the questionnaire.

[42]The identification numbers of these subjects are highlighted in *Online Appendix C*. The majority of them present an inverted U-shaped payback pattern, which yields estimated psychological types $(\hat{G}, \hat{R})$ equally distributed across the following three categories: guilt prevails, reciprocity prevails, and balanced motivations. Although such categorizations according to $(\hat{G}, \hat{R})$ are "forced," the answers of these subjects to the debriefing questions—available in *Online Appendix C*—seem to confirm that the categorization makes sense.

Categories of elicited psychological types	Estimated payback function	Treatment			
		NoQ	*QnoD*	*NoQ-QnoD*	*QD*
Guilt prevails ($\hat{G} > \hat{R}$, $\hat{R}$ small)	$\hat{\xi}'(\alpha) > 0$	23 (1)	20 (2)	43 (3)	45 (4)
Guilt prevails for high α ($\hat{G} > \hat{R}$, $\hat{R}$ not small)	$\hat{\xi}(\alpha)$ U-shaped	3 (0)	2 (0)	5 (0)	3 (0)
Reciprocity prevails ($\hat{G} < \hat{R}$)	$\hat{\xi}'(\alpha) < 0$	7 (2)	7 (4)	14 (6)	12 (2)
Balanced motivations ($\hat{G} = \hat{R}$)	$\hat{\xi}'(\alpha) = 0$	3 (2)	2 (1)	5 (3)	9 (3)
Selfish preferences ($\hat{G} = \hat{R} = 0$)	$\hat{\xi}(\alpha) = 0$	4 (0)	9 (0)	13 (0)	11 (1)
TOTAL		40 (5)	40 (7)	80 (12)	80 (10)

Table 4 Categorization of B-subjects according to the payback pattern.

The table reports, for each treatment and category of psychological types: the number of B-subjects with elicited $(\hat{G}, \hat{R})$ in that category; within parentheses, the number of B-subjects with elicited $(\hat{G}, \hat{R})$ in that category, but with payback pattern not captured by the corresponding shape of $\xi(\alpha)$. Column *NoQ-QnoD* pools the observations of *NoQ* and *QnoD*.

Together with the right panel of Figure 3, Table 4 also shows that, independently of the treatment, the guilt component is prevalent for more than half of the B-subjects, while reciprocity prevails for only 16% of them.[43] There is also a non-negligible number of B-subjects (5%) for whom guilt prevails when A's first-order belief is high, and reciprocity prevails otherwise (U-shaped payback function). The remaining subjects have a flat estimated payback function (balanced motivations). The majority of them are selfish (0 payback regardless of α, 15% of the sample). The estimated payback function of the others (9% of the sample) is consistent with inequity aversion: These subjects aim at an interior distribution independent of α.

The following statement summarizes the main experimental findings about the distribution of B-subjects' payback patterns.

[43] In a trust game similar to the one of Charness and Dufwenberg (2006), Ederer & Stremitzer (2017) find that more than half of the trustees exhibit guilt aversion. Bellemare *et al.* (2018), using an elicitation method similar to ours, also find that the majority of trustees are guilt averse (see Menu treatment of Experiment 1, p. 237). None of these studies investigate trustees' recipocity.

Result 1 The great majority (86%) of B-subjects' payback patterns are consistent with the theoretical shapes implied by our model. Across all B-subjects, the estimated payback functions $\hat{\xi}(\alpha)$ are mostly *belief-dependent* (76%); of these, the guilt component is prevalent for 72%, while reciprocity prevails for only 21%. Similar results hold for the subpopulations of subjects within the different treatments.

4.2 Behavior under disclosure of the filled-in questionnaire

This subsection is split into two parts. First, we organize B-subjects and matched A-subjects according to the complete-information predictions using the estimated psychological type $(\hat{G}, \hat{R})$ obtained from B-subjects' payback pattern (predicted behavior). Second, we compare observed behavior with predicted behavior, at the pair and individual level.

Figure 4 reports the *observed vs. predicted behavior* of matched *pairs* in phase 3 of *QD*, the only phase in our experimental design that supposedly approximates a Trust Minigame with complete information. Figure 4a refers to the three regions of the parameter space (G, R) of Figure 1, which correspond to the complete-information predictions of rationalizability based on forward induction of Proposition 2. Figure 4b refers to the two regions of the parameter space (G, R) of Figure 2, which correspond to the equilibrium refinement of Proposition 3. The latter extends the FI-dominance regions $\mathbb{T}$ and $\mathbb{S}$ of Figure 1 to each $(G, R) \in \mathbb{R}^2$.

In both figures, for each region and for each category of psychological type from Table 4 (guilt prevails, reciprocity prevails, etc.), we report in bold the number of "**classified**" B-subjects and in Italics the number of remaining ("*unclassified*") subjects. Classified B-subjects in Figure 4a have a $(\hat{G}, \hat{R})$ that can be assigned to one of the *three* regions of the parameter space (G, R) of Figure 1 with a level of significance of at most 10% (P-values estimated by bootstrap). For Figure 4b the same holds for the *two* regions of the parameter space (G, R) of Figure 2.

In each figure, an estimated psychological type $(\hat{G}, \hat{R})$ in the light-grey region $\mathbb{T}$ leads to a prediction of (*Dissolve,Take*) for the corresponding matched pair, while the dark-grey region $\mathbb{S}$ refers to a prediction of (*Continue,Share*). Therefore, we call "**predictable**" the classified B-subjects such that $(\hat{G}, \hat{R}) \in \mathbb{S} \cup \mathbb{T}$. Conversely, classified B-subjects with $(\hat{G}, \hat{R}) \in (\mathbb{S} \cup \mathbb{T})^c$ (white-colored region of Figure 4a) are not predictable, since any strategy profile of the corresponding matched pair is rationalizable. Before the number of predictable B-subjects in *QD* (bold font) we report the number of the corresponding matched pairs who behave as predicted in phase 3 of *QD* (normal font).

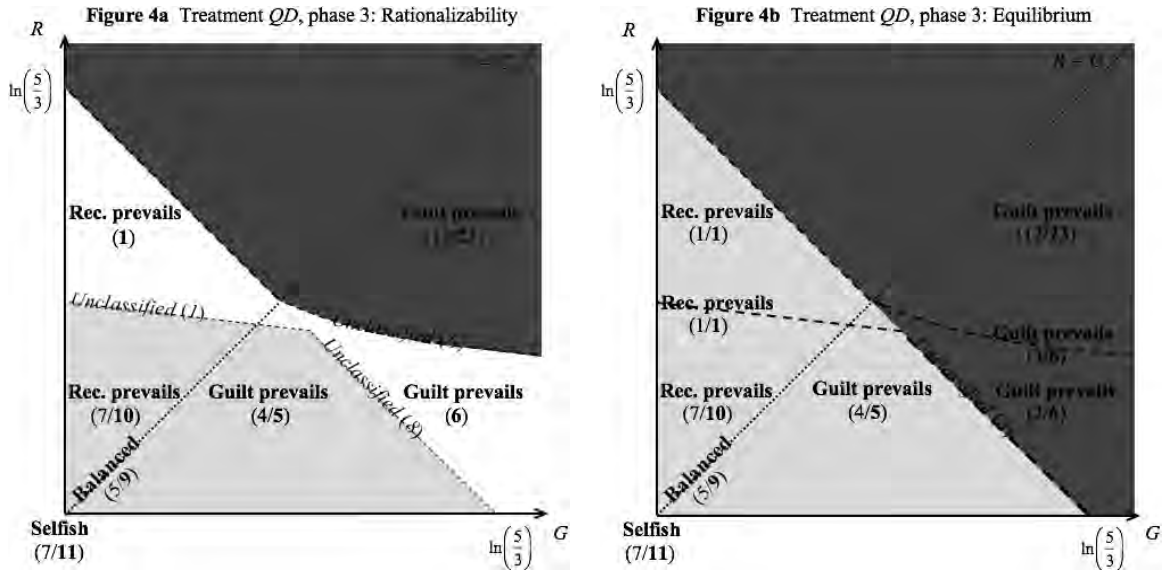

Figure 4 Observed *vs.* predicted behavior (strategy pairs) in phase 3 of *QD*.

Figure 4a refers to the complete-information predictions of *rationalizability* (Proposition 2). Figure 4b refers to the complete-information *equilibrium* predictions (Proposition 3). Each ratio indicates observed (normal font) *vs.* predicted (bold) behavior in phase 3 of *QD*. Number in Italics indicate unclassified B-subjects. Estimated types in the white (intermediate) region of Figure 4a are classified, but do not yield a prediction according to rationalizability; thus, we do not report observed behavior.

Predicted behavior of A-B pairs Given the estimated psychological type $(\hat{G}, \hat{R})$, we can make a prediction for about 73% (58/80) of pairs in phase 3 of *QD* according to rationalizability (bold numbers in regions $\mathbb{T}$ and $\mathbb{S}$ of Figure 4a) and for 90% (72/80) of pairs according to the equilibrium predictions (bold numbers in Figure 4b). For the latter, the (*Dissolve*,*Take*) region includes all pairs with a B-subject for whom reciprocity prevails, all pairs with a "balanced" B-subject, and, obviously, all selfish B-subjects. Conversely, for the great majority of pairs with a B-subject for whom guilt prevails, (*Continue*,*Share*) is the complete-information prediction.

In particular, for *all* B-subjects in the (*Continue*,*Share*) region of Figure 4b, we find that $\hat{G} > \hat{R}$ and, for all but one of these subjects, we have $\hat{G} > \ln(5/3)$, i.e., higher than the theoretical threshold for (*Continue*,*Share*) in Proposition 3. Hence guilt aversion is, by itself, high enough to yield the cooperative equilibrium under complete information. For this reason, from now on, we refer to the B-subjects in the (*Continue*,*Share*) region of Figure 4b as **"high-guilt" types** (35/72) and to the B-subjects in the (*Dissolve*, *Take*) region of Figure 4b as **"low-guilt" types** (37/72). The latter subgroup includes all predictable B-subjects with $\hat{R} > \hat{G}$ and all those with $\hat{G} = \hat{R} \geq 0$.

The following result summarizes the main experimental findings about B-subjects' predicted behavior under complete information.

Result 2 Given the estimated guilt and reciprocity components, all B-subjects predicted to choose *Share* under complete information are "high-guilt" types.

This further corroborates the hypothesis that guilt aversion is the main driver of cooperation in the Trust Minigame.

Observed behavior of A-B pairs In Figure 4a pooled ratios of observed *vs.* predicted behavior in phase 3 of *QD* show a 60% (35/58) rate of success of the complete-information predictions for phase 3 of *QD*. The rate of success is not significantly different in Figure 4b (58%, 42/72; χ^2 test, *P-value* $= 0.816$), where, due to equilibrium refinement, we also consider as predictable the 7 intermediate types and 7 unclassified types of Figure 4a. Both rates of success are significantly higher than the one (25%) of a random guess over the four possible strategy profiles (χ^2 test, *P-value* $= 0.000$ for both Figure 4a and Figure 4b).

Our complete-information predictions are particularly successful for pairs predicted to choose (*Dissolve*,*Take*): 66% (23/35) in Figure 4a and 68% (25/37) in Figure 4b. They are slightly less successful for pairs predicted to choose (*Continue*,*Share*): 52% (12/23) in Figure 4a and 49% (17/35) in Figure 4b (both rates of success are still significantly higher—at the 1% level—than the one of a random guess). This can be due to lower than predicted conditional second-order beliefs of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ (see the last paragraph of Section 5 for a possible explanation).

The following result summarizes the main experimental findings about behavior and beliefs of matched pairs under complete information.

Result 3 Complete-information *rationalizability* explains 60% of the observed behavior of predicted matched pairs after questionnaire disclosure (phase 3 of treatment *QD*). In particular, 66% of pairs in the (*Dissolve*,*Take*) region and 52% of pairs in the (*Continue*,*Share*) region behave as predicted. Similar results are found under the complete-information *equilibrium* refinement.

In Figure 5 we deepen the analysis presented in Figure 4b: We present *subjects' observed choices and beliefs* in phase 3 of *QD*, disentangled by role and by B's psychological type. Despite a slightly lower rate of success, we consider equilibrium rather than rationalizability predictions since, by construction, they capture a higher number of pairs in the two regions of predictions (*Continue*,*Share*) and (*Dissolve*,*Take*): 35 subjects in the former region (high-guilt types) and 37 subjects in the latter region (low-guilt types) of Figure 4b. We have

checked that all the results below also hold if we rely on the rationalizability predictions of Figure 4a.

First, we discuss experimental results about *choices and first-order beliefs* of A-subjects in phase 3 of *QD*.[44] Then, we discuss experimental results about *choices*, first *and second-order beliefs* of B-subjects in phase 3 of *QD*.[45]

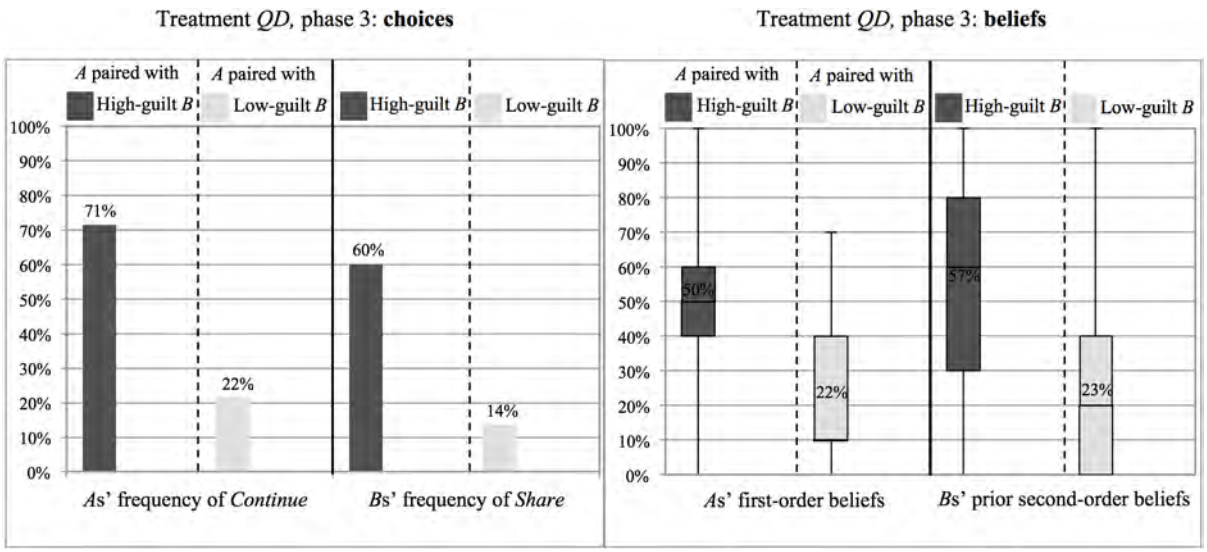

Figure 5 A's and B's choices and beliefs in phase 3 of *QD*, disentangled by B's type.

The figure reports, for phase 3 of *QD*: on the left panel, the frequency of A-subjects' *Continue* choices and of matched B-subjects' *Share* choices; on the right panel, the box plot and average of A-subjects' first-order belief and B-subjects' unconditional second-order belief of *Share*. The color code is related to Figure 4b: all high-guilt Bs belong to the dark-grey (*Continue*,*Share*) region, all low-guilt Bs belong to the light-grey (*Dissolve*, *Take*) region.

[44] A clarification about A-subjects' first-order beliefs in Figure 5 is in order. Recall (see Section 2.2) that A's elicited first-order belief is not only about the matched B, but about all the 10 B-subjects in the session; hence, according to our complete-information predictions, we should get an elicited α that is less polarized than the true one. For example, an A who faces a high-guilt B in phase 3 of *QD* is asked how many of the 10 B-subjects in the session (the matched B and the other nine) will *Share*, and she can rationally presume—despite the disclosed filled-in questionnaire of the matched high-guilt B—that there are some low-guilt B-subjects in the session.

[45] A clarification about B-subjects' second-order beliefs in Figure 5 is in order. B knows that A did not state a belief solely about the choice of the matched B (see previous footnote); hence, according to our complete-information predictions, we should get an elicited unconditional second-order belief, $\mathbb{E}_B(\widetilde{\alpha})$, that is less polarized than the true one (Charness & Dufwenberg 2006 face a similar problem: see their footnote 12). Furthermore, recall that we elicit the unconditional, not the conditional second-order belief. The latter is the relevant correlate of B's propensity to share, but the former shows how B thinks about the game. For example, when B is a low-guilt type, knowing that A observes this, he should expect that α is very low. But B's conditional expectation of α given *Continue* can reasonably be larger than 1/2. With the same caveat explained above (B knows that A's stated belief is not just about him), the elicited unconditional belief $\mathbb{E}_B(\widetilde{\alpha})$ is a rough estimate of the conditional belief β for those B-subjects indicating *Continue* as the most likely choice of the matched A-subject.

Behavior and beliefs of *A*-subjects As reported in Figure 5, A-subjects matched with a high-guilt B-subject show a significantly higher (at the 1% level) frequency of *Continue* (+50%, χ^2 test) and first-order belief α (+29% on average, Mann-Whitney test). A significant (at the 1% level) positive correlation is found between the *Continue* choice and α (rank-biserial correlation coefficient, **Somers'** $D = 0.59$).

A further result supporting the complete-information predictions is the significant (at the 1% level) positive correlation found in phase 3 of *QD* between $(\hat{G} + \hat{R})$—the feature of B's estimated psychological type $(\hat{G}, \hat{R})$ relevant for the equilibrium analysis of Proposition 3—and both A's choice of *Continue* ($D = 0.52$) and α (**Spearman**'s correlation coefficient $\rho = 0.44$). This is mainly due to the guilt component $\hat{G}$ ($D = 0.54$ with *Continue*, $\rho = 0.48$ with α), while for $\hat{R}$ we find a low negative correlation with both A-subjects' choice ($D = -0.20$, *P-value* $= 0.145$) and belief ($\rho = -0.26$, *P-value* $= 0.025$).[46]

Finally, if we disentangle the A-subjects in phase 3 of *QD* according to the matched (estimated) psychological type—high-guilt *vs.* low-guilt—and we focus on any of the two subgroups separately, we find no significant correlation between $\hat{G}+\hat{R}$ and both the *Continue* choice and α: the largest (in absolute value) of the 4 correlation coefficients is 0.14 (*P-value* $= 0.550$) between $\hat{G} + \hat{R}$ and *Continue* for As matched with low-guilt Bs. This is in line with the complete-information predictions of Proposition 3 given the matched B's elicited psychological types.

The following result summarizes the more salient experimental findings about A-subjects' behavior and beliefs under complete information.

Result 4 In line with the complete-information predictions, after questionnaire disclosure, both the frequency of *Continue* choices and the first-order beliefs are significantly higher for A-subjects matched with high-guilt B-subjects. More generally, both the propensity to *Continue* and A's first-order beliefs are positively correlated with the disclosed guilt type of B.

Behavior and beliefs of *B*-subjects As reported in Figure 5, high-guilt B-subjects show a significantly higher (at the 1% level) frequency of *Share* (+46%, χ^2 test) and unconditional second-order beliefs $\mathbb{E}_B[\widetilde{\alpha}]$ (+34% on average, Mann-Whitney test) than low-guilt B-subjects.

As for B-subjects' first-order beliefs, recall that we only ask them whether they expect *Continue* or *Dissolve*, i.e., a (coarse) feature of their first-order beliefs. For ease of notation,

[46] We verified that $\hat{G}$ and $\hat{R}$ are statistically independent ($\rho = -0.10$, *P-value* $= 0.110$). This allows us to run the correlation analysis with A's choice and first-order belief for $\hat{G}$ and $\hat{R}$ separately. Furthermore, the controls for no correlation between $\hat{p}$ and both A's choice of *Continue* ($D = 0.08$, *P-value* $= 0.538$) and α ($\rho = -0.08$, *P-value* $= 0.485$) work as they should.

and with an abuse of language, we refer to such reported beliefs as B-subjects' **first-order point-belief**. With this, we find a strongly significant positive correlation between *Share* and the first-order point-belief ($\rho = 0.44$, *P-value* $= 0.000$).

The correlation between *Share* and $\mathbb{E}_B[\widetilde{\alpha}]$ is again strongly significant ($D = 0.65$, *P-value* $= 0.000$). We find the same significant correlation if we consider only B-subjects for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of the *conditional* second-order belief β (those with *Continue* as first-order point-belief). Focusing on the latter subjects, we observe that 90% (19/21) of those classified as high-guilt types and with $\mathbb{E}_B[\widetilde{\alpha}] \geq 1/2$ choose *Share*.

Results about the positive correlation (significant at the 1% level) between $\hat{G} + \hat{R}$ and, respectively, B's *Share* choice ($D = 0.52$), first-order point belief ($r_{pb} = 0.59$), and $\mathbb{E}_B[\widetilde{\alpha}]$ ($\rho = 0.46$) are consistent with the theoretical predictions. In particular, since B is aware that his psychological type $(\hat{G}, \hat{R})$ is disclosed to A, his beliefs about A's behavior and beliefs move with $\hat{G} + \hat{R}$.[47]

Disentangling by psychological type—high-guilt *vs.* low-guilt—, we find no significant correlation between $\hat{G} + \hat{R}$ and B-subjects' choices and first and second-order beliefs, in any of the two subgroups considered separately: the largest (in absolute value) of the 8 correlation coefficients is 0.15 (*P-value* $= 0.508$) between $\hat{G} + \hat{R}$ and the first-order point-belief of *Share* for low-guilt Bs. This confirms the complete-information predictions: B's choice depends on whether $G + R$ is above or below the threshold in Proposition 3, but not on their precise value.

The following result summarizes the more salient experimental findings about B-subjects' behavior and beliefs under complete information.

Result 5 In line with the complete-information predictions, after questionnaire disclosure the frequency of *Share* choices, the first- and the second-order unconditional beliefs are significantly higher for high-guilt than for low-guilt B-subjects. More generally, cooperation and B's first- and second-order unconditional beliefs are positively correlated with the estimated guilt type of B.

4.3 Behavior without disclosure of the filled-in questionnaire

In this section, we focus on the **"incomplete-information phases,"** i.e., those phase-treatment combinations where the filled-in questionnaire is not disclosed (phase 1 of *QD*, phases 1 and 3 of *NoQ-QnoD*). In these phases, subjects play a Trust Minigame with incomplete information about B's psychological type. Throughout this subsection, we provide

[47] As for A-subjects, also for B-subjects we find a significant (at the 1% level) positive correlation of choices and beliefs with $\hat{G}$, and a low negative (non-significant) correlation with $\hat{R}$. As for elicited ex-post feeling mitigation, no correlation between $\hat{p}$ and, respectively, B's *Share* choice ($D = 0.08$, *P-value* $= 0.557$), first-order point belief ($D = 0.14$, *P-value* $= 0.295$), and $\mathbb{E}_B(\widetilde{\alpha})$ ($\rho = -0.01$, *P-value* $= 0.938$) is found.

aggregate results about the incomplete-information phases, because we do not find significant between-treatment, or within-treatment differences. In particular, due to a significant correlation in subjects' choices and beliefs across phase 1 and phase 3 of *NoQ-QnoD*, we only consider phase 3 of this treatment, which is relevant for between-treatment comparison with phase 3 of *QD* (see Section 4.4). Therefore, all the results in this subsection rely on pooled data of phase 1 of *QD* and of phase 3 of *NoQ-QnoD*. We have checked that all results below hold if considering data of phase 1 rather than phase 3 of *NoQ-QnoD*.[48]

To check our auxiliary assumptions and theoretical predictions about subjects' behavior and beliefs, we analyze the experimental results in light of the qualitative features of the non-degenerate equilibrium described in Proposition 5 (whose statement (4) incorporates the qualitative features of the incomplete-information rationalizability predictions of Proposition 4):

(1) A-heterogeneity A-subjects' first-order beliefs are heterogeneous and dispersed: Only 23% (1%) of A-subjects have $\alpha = 0$ ($\alpha = 1$), the coefficient of variation of α is 0.89. We also find a significant difference (at the 1% level) in the frequency of *Continue* choices (81% *vs.* 14%) between A-subjects with $\alpha \geq 1/2$ and A-subjects with $\alpha < 1/2$. This result corroborates the assumption that A has selfish risk-neutral preferences (hence she should choose *Continue* if and only if $\alpha \geq 1/2$).

(2) B-heterogeneity B-subjects have heterogeneous first-order point-beliefs about A's strategies, with 41% (59%) of B-subjects reporting *Continue* (*Dissolve*). The unconditional second-order beliefs are heterogeneous and dispersed: Only 26% (4%) of B-subjects have $\mathbb{E}_B[\widetilde{\alpha}] = 0$ ($\mathbb{E}_B[\widetilde{\alpha}] = 1$), the coefficient of variation of $\mathbb{E}_B[\widetilde{\alpha}]$ is 0.90. If we consider only B-subjects for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of β (i.e., those whose first-order point-belief is *Continue*), we find that almost all of them (94%) have $\mathbb{E}_B[\widetilde{\alpha}] > 0$, but only 43% of them have $\mathbb{E}_B[\widetilde{\alpha}] \geq 1/2$.

(3.i) Independence between roles As expected in a random-matching setting, we find that A's choice is independent of the matched B's choice ($\rho = -0.02$), $\hat{G} + \hat{R}$ ($D = -0.02$), first-order point-belief ($\rho = 0.04$), and $\mathbb{E}_B[\widetilde{\alpha}]$ ($D = 0.05$). A similar result holds for A's first-order belief (low correlation $D = -0.20$ at a 10% level with B's choice, $\rho = -0.02$ with $\hat{G} + \hat{R}$, $D = 0.01$ with B's first-order point-belief and $\rho = 0.02$ with $\mathbb{E}_B[\widetilde{\alpha}]$).

[48] Compare the right panel of Figure 6 below—observed *vs.* predicted behavior in phase 3 of *NoQ-QnoD*—to Figure C.2 in *Online Appendix C*. Figure C.2 reports observed *vs.* predicted behavior in phase 1 of *NoQ-QnoD*.

(3.ii) Independence within roles Second-order beliefs of B are independent of $\hat{G}+\hat{R}$ ($\rho = 0.07$, *P-value* $= 0.363$), and first-order point-belief shows a low positive correlation with $\hat{G}+\hat{R}$ ($D = 0.22$, *P-value* $= 0.016$). This corroborates our auxiliary assumption that the epistemic component of B's type is independent of the psychological component.

(4) FI-dominance We organize B-subjects' choices according to the incomplete-information predictions of Proposition 4 using the estimated parameters $\hat{G}$ and $\hat{R}$ obtained from the pay-back pattern. Figure 6 refers to the three regions of predictions in the parameter space (G, R) of Figure 1. It has been constructed using the same method and notation as Figure 4a, although the latter refers to predicted behavior of *A-B pairs* under questionnaire disclosure. Absent questionnaire disclosure, Figure 6 only refers to B-subjects: On the left panel, we report B-subjects' observed (normal font) *vs.* predicted (bold) behavior in phase 1 of *QD*; on the right panel, we report observed *vs.* predicted behavior in phase 3 of *NoQ-QnoD*.

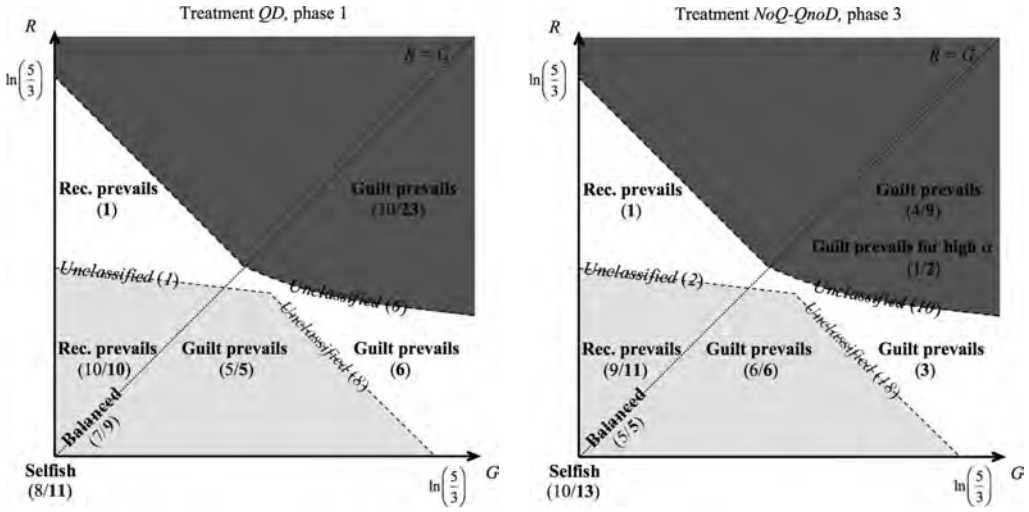

Figure 6 Observed *vs.* predicted behavior of B-subjects in the incomplete-information phases. The figure refers to the three regions of the parameter space (G,R) of Figure 1. The classification method and notation are the same as in Figure 4a.

In *QD*, we are able to classify 65/80 B-subjects,[49] and 58 out of 65 are predictable, i.e. with $(\hat{G}, \hat{R}) \in \mathbb{S} \cup \mathbb{T}$ (by construction, these numbers are the same as in Figure 4a); in *NoQ-QnoD*, we classify 50/80 B-subjects, and 46 out of 50 are predictable. Relying on the incomplete-information predictions in Figure 5, and considering together phase 1 of *QD* (left panel) and phase 3 of *NoQ-QnoD* (right panel), we find that *Share* is chosen by 44% of

[49] As for Figure 4a, classified B-subjects in Figure 5 are those for whom $(\hat{G}, \hat{R})$ can be assigned to one of the three regions of Figure 1 with a level of significance of at most 10% (P-values estimated by bootstrap).

B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ (dark-grey region), while it is chosen by 14% of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{T}$ (light-grey region), the difference is significant at the 1% level. The fact that less than half of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ choose *Share* seems to be mostly explained by a failure of the forward-induction inference that $\beta \geq 1/2$ (see the test of statement (2) above). Indeed, if we consider only B-subjects for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of β (first-order point-belief *Continue*), we find that 88% of those with $(\hat{G}, \hat{R}) \in \mathbb{S}$ and $\mathbb{E}_B[\widetilde{\alpha}] \geq 1/2$ choose *Share*. We further discuss this issue in the last paragraph of Section 5.

(5) Choice-belief correlation We find a significant positive correlation ($D = 0.35$, *P-value* $= 0.057$) between *Share* and $\mathbb{E}_B[\widetilde{\alpha}]$ for B-subjects with $\hat{G} \geq 2\hat{R}$ and for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of β (i.e., those whose first-order point-belief is *Continue*).

The following result summarizes the more salient experimental findings about behavior and beliefs under incomplete information.

Result 6 In line with the incomplete-information predictions, in the phase-treatment combinations where the questionnaire is not disclosed, we find heterogeneous and dispersed beliefs about B's strategy, about A's strategy, and about the elicited α. For B-subjects with $\hat{G} \geq 2\hat{R}$ who expect *Continue*, the *Share* choice is positively correlated with the belief about α. Furthermore, *Share* is chosen by only 14% of B-subjects predicted to choose *Take* (i.e., with $(\hat{G}, \hat{R}) \in \mathbb{T}$); this fraction is significantly higher for B-subjects predicted to choose *Share* (i.e., with $(\hat{G}, \hat{R}) \in \mathbb{S}$), although it is only 44%.

4.4 Disclosure *vs.* non-disclosure of the filled-in questionnaire

We conclude the data analysis with a qualitative comparison of behavior and beliefs under complete *vs.* incomplete information. As anticipated in Section 3.4, we rely on the equilibrium predictions of Proposition 3. In particular, we rely on the separation between high-guilt and low-guilt types introduced in the complete-information case—see Figure 4b for the distribution of estimated psychological types in treatment *QD*. As expected, the ratio of high *vs.* low-guilt B-subjects in the aggregate treatment *NoQ-QnoD* (24/38) is not significantly different from the *QD* treatment (35/37, *P-value* $= 0.250$; χ^2 test).[50]

With this, we extend the analysis in Section 4.2. First, we compare frequencies of strategy profiles chosen by complete-information predictable *pairs* in phase 3 of *QD* *vs.* the incomplete-information phases relevant for within-treatment and between-treatment comparisons (respectively, phase 1 of *QD* and phase 3 of *NoQ-QnoD*).

[50] We replicated the exercise leading to Figure 4b also for the 80 B-subjects in *NoQ-QnoD* (see Figure C.1 in *Online Appendix C*). In particular, we find that all B-subjects in the (*Continue*,*Share*) region (24/80) are high-guilt types ($\hat{G} > \hat{R}$).

Then we analyze subjects' choices and beliefs—disentangled by *role* and by B's type—to make within-treatment comparisons (phase 1 *vs.* phase 3 of *QD*) and between-treatment comparisons (phase 3 of *QD vs.* phase 3 of *NoQ-QnoD*).[51]

Observed behavior of A-B pairs Table 5 reports the frequencies of strategy profiles for *all* the complete-information predictable pairs, comparing phase 3 of *QD* with the incomplete-information phases 1 of *QD* and 3 of *NoQ-QnoD*. In line with the complete-information predictions, in phase 3 of *QD* there is a significant correlation ($\rho = 0.35$, *P-value* $= 0.002$) between *Continue* (resp. *Dissolve*) and *Share* (resp. *Take*); a significant correlation ($\rho = 0.33$, *P-value* $= 0.005$) is also found between the elicited values of α and $\mathbb{E}_B[\widetilde{\alpha}]$. In both phase 1 of *QD* and phase 3 of *NoQ-QnoD*, as expected in a random-matching setting, the choices of A and B are independent ($\rho = -0.02$ on pooled data, *P-value* $= 0.775$) and the same holds for their beliefs ($\rho = 0.02$ on pooled data, *P-value* $= 0.819$).

Phase 1

	Take	*Share*	
Diss	47%	18%	65%
Cont	24%	11%	35%
	71%	29%	

Phase 3

	Take	*Share*	
Diss	43%	11%	54%
Cont	21%	25%	46%
	64%	36%	

QD

	Take	*Share*	
Diss	64%	19%	83%
Cont	15%	2%	17%
	79%	21%	

NoQ-QnoD

Table 5 Frequencies of observed strategy profiles, disentangled by phase-treatment combination.

In other words, under disclosure there is a polarization of choices and beliefs along an "axis of trust," that is—referring to the phase 3-*QD* matrix in Table 5—, a higher concentration of pairs on the main diagonal (low trust in the top-left box, high trust in the bottom-right box) compared to the product of the marginals. In line with our theoretical predictions (Section 3.4), this polarization is not present without disclosure. We interpret this result as follows: The correlation found between *Continue* (resp., *Dissolve*) and *Share* (resp., *Take*) and between α and $\mathbb{E}_B[\widetilde{\alpha}]$ in phase 3 of *QD* is due to the disclosure of B's psychological type to A. When A receives the filled-in questionnaire of a high-guilt (resp., low-guilt) type,

[51] We implemented a stranger-matching design: in each treatment, As and Bs are randomly re-matched so as to have different pairs in phase 1 and in phase 3 and avoid repeated-game effects. However, with the goal of providing a clean check of within-treatment differences, throughout this subsection we analyze pairs' behavior in phase 1 of each treatment according to the matching of phase 3. This can be done at no cost, since A's (B's) choice in phase 1 is told to the matched B (A) only at the end of the experiment.

she tends to believe that B would choose *Share* (resp., *Take*), hence she tends to choose *Continue* (resp., *Dissolve*). Knowing this, a disclosed high-guilt (resp., low-guilt) type tends to choose *Share* (resp., *Take*).

In Figures 7 and 8 we deepen the analysis presented in Table 5. With the complete-information predictions of Proposition 3 in mind, we extend Figure 5, which only refers to phase 3 of *QD*. Figure 7 shows the within-treatment comparisons of choices (frequencies) and beliefs (average and box plot) disentangled by estimated psychological type of B (high *vs.* low-guilt). Figure 8 shows the analogous between-treatment comparisons.

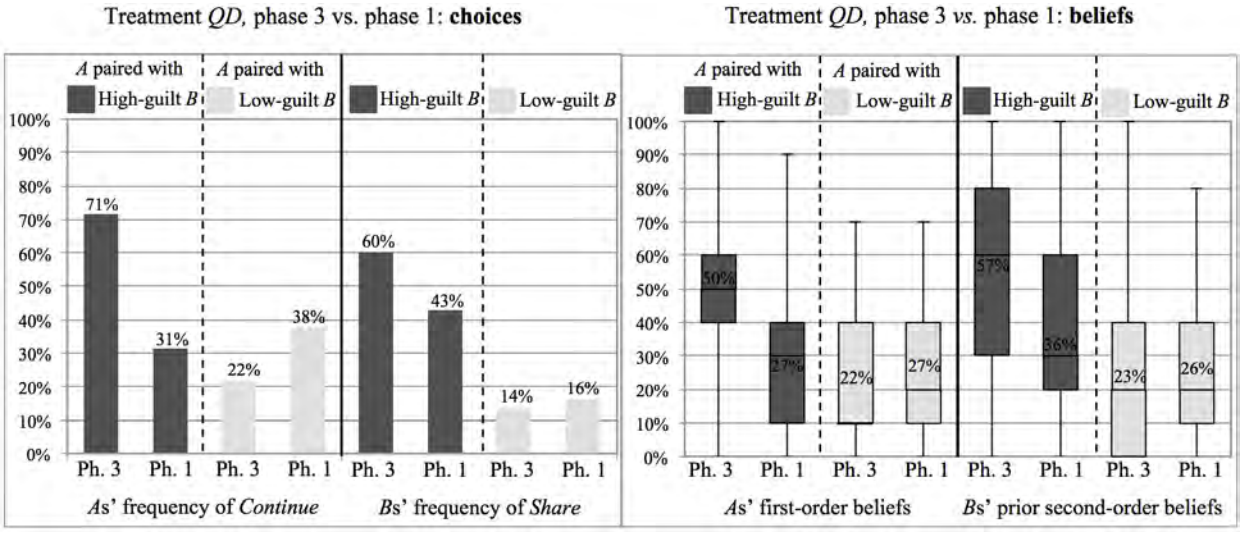

Figure 7 A's and B's choices and beliefs in phase 3 *vs.* phase 1 of *QD*, disentangled by B's type.

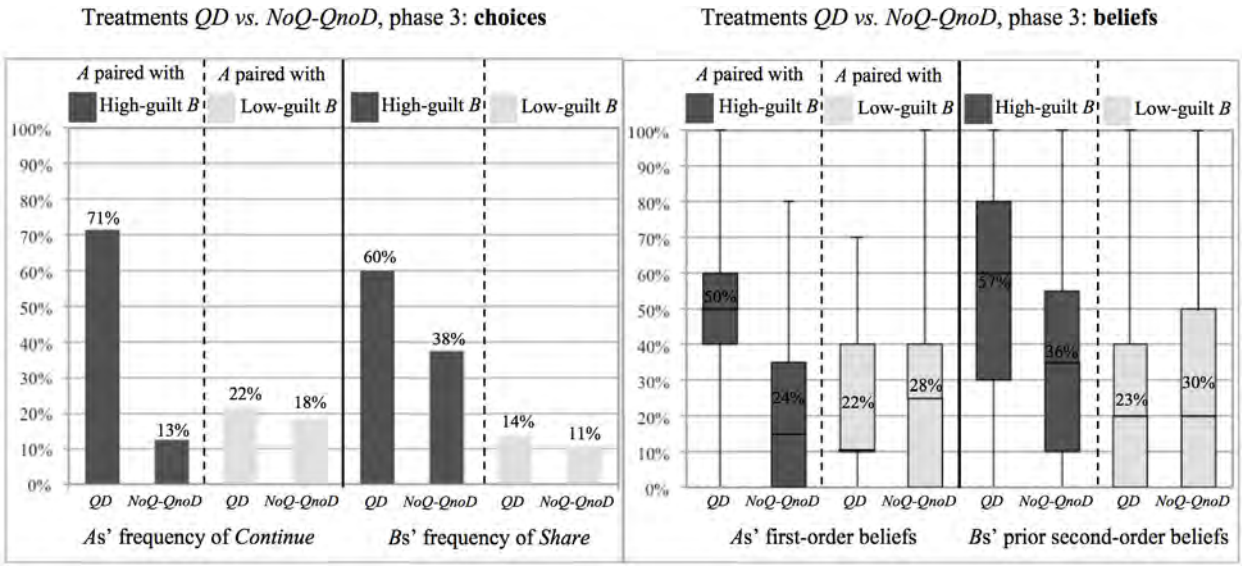

Figure 8 A's and B's choices and beliefs in phase 3 of *QD vs. NoQ-QnoD*, by B's type.

Behavior and beliefs of *A*-subjects The controls for A-subjects work as they should: In each incomplete-information phase, we find no significant difference in the frequency of *Continue* and in the distribution of the first-order beliefs between A-subjects matched with a high-guilt type and A-subjects matched with a low-guilt one.[52]

Between-treatment and within-treatment comparisons work very well for A-subjects matched with a *high-guilt* type: Between treatments, we find a significantly (at the 1% level) higher frequency of *Continue* (+59%, χ^2 test) and α (+26% on average, Mann-Whitney test) in phase 3 of *QD* than in phase 3 of *NoQ-QnoD*. Within treatment, we find a similar result by comparing phase 3 to phase 1 of *QD*: respectively, +40% and +23% on average, both significant at 1%. A Wilcoxon matched-pairs signed-ranks test confirms the latter result: moving from phase 1 to phase 3 of *QD*, 17/35 (*P-value* = 0.002) A-subjects matched with a *high-guilt* type switched from *Dissolve* to *Continue* (only 3/35 switched from *Continue* to *Dissolve*), and 26/35 increased (4/35 decreased, *P-value* = 0.000) their α.

Between-treatment and within-treatment comparisons are less striking for A-subjects matched with a *low-guilt* type: No significant difference is found (+3% for *Continue* and −6% for α) by comparing phase 3 between *QD* and *NoQ-QnoD*. The decrease from phase 1 to phase 3 of the frequency of *Continue* (−16%) and of α (−5%) within *QD* is not significant, although the ratio of A-subjects switching from *Continue* to *Dissolve* is higher than the ratio of those switching from *Dissolve* to *Continue* (10/37 *vs.* 4/37, signed-ranks test, *P-value* = 0.109), and 17/37 decreased *vs.* 13/37 increased α (signed-ranks test, *P-value* = 0.338).

Behavior and beliefs of *B*-subjects Consistently with our hypothesis (see Section 3.4), *guilt prevails in FI-underdetermined subjects*: the majority of subjects in $(\mathbb{S} \cup \mathbb{T})^c$ display high guilt (12/14 in *QD* and 13/16 In *NoQ-QnoD*).

By Proposition 5, in each incomplete-information phase we expect high-guilt Bs to show a significantly higher frequency of *Share*, but not significantly different second-order beliefs than low-guilt ones; the latter prediction is due to the assumption of independence between the psychological and epistemic component of B's type. In Section 4.3 we have already shown that both predictions hold (see, respectively, the test of statements (4) and (3.ii) of Proposition 5). This can be also detected by looking at the histograms (for statement (4)) and box plots (for statement (3.ii)) of high-guilt *vs.* low-guilt Bs in phase 1 of *QD* (Figure 7) and phase 3 of *NoQ-QnoD* (Figure 8).

For phase 3 of *QD*, between-treatment and within-treatment comparisons work quite well for *high-guilt* B-subjects: Between treatments, we find a higher frequency of *Share* (+23%,

[52]For phase 1 of *NoQ-QnoD*, this is shown in Figure C.3 in *Online Appendix C*, where we report As and Bs' choices and beliefs in phase 1 of *QD vs. NoQ-QnoD*, disentangled by B's type (high-guilt *vs.* low-guilt). Figure C.3 shows no significant difference in phase 1 of *QD vs. NoQ-QnoD*: for As' choices and beliefs independently of the matched Bs' type; for Bs' choices and beliefs given his type.

χ^2 test, *P-value* $= 0.089$) and significantly higher second-order beliefs (+21% on average, Mann-Whitney test, *P-value* $= 0.012$) by comparing phase 3 of *QD* to phase 3 of *NoQ-QnoD*. Within treatment, we find similar differences by comparing phase 3 to phase 1 of *QD*: +17% (*P-value* $= 0.151$) for the frequency of *Share*, and +20% on average (*P-value* $= 0.005$) for $\mathbb{E}_B[\widetilde{\alpha}]$. A Wilcoxon matched-pairs signed-ranks test confirms the non-significant difference for choices (12/35 *vs.* 6/35, *P-value* $= 0.157$) and the significant difference for unconditional second-order beliefs (25/35 *vs.* 9/35, *P-value* $= 0.005$).

The within-subject difference in the frequency of *Share* is smaller than the between-subject one. This slight "difference in difference" observed in our data is due to random differences in the distribution of psychological types in the two treatments: According to our equilibrium predictions, the only psychological types whose behavior is expected to change with the information regime are those outside the FI-dominance regions of Figure 1 (white-colored region $(\mathbb{S} \cup \mathbb{T})^c$). The fraction of high-guilt types with $(\hat{G}, \hat{R}) \in (\mathbb{S} \cup \mathbb{T})^c$ relative to the predictable ones is lower in *QD* (13/35, left panel of Figure 6) than in *NoQ-QnoD* (13/24, right panel of Figure 6). This explains the smaller difference for the within-subject comparison.

Between-treatment and within-treatment comparisons work well also for *low-guilt* *B*-subjects: The predicted behavior is the same under both complete and incomplete information (35/37 *B*-subjects in the (*Dissolve*, *Take*) region of Figure 4b also have $(\hat{G}, \hat{R}) \in \mathbb{T}$ in the left panel of Figure 6), and indeed we find no significant difference in the frequency of *Take*. Furthermore, as predicted, $\mathbb{E}_B[\widetilde{\alpha}]$ is lower in phase 3 of *QD*, although not significantly. All this holds regardless of whether we compare phase 3 between *QD* and *NoQ-QnoD*, or phase 3 to phase 1 within *QD* (Wilcoxon matched-pairs signed-ranks test: *P-value* $= 0.763$ for choices, and *P-value* $= 0.393$ for $\mathbb{E}_B[\widetilde{\alpha}]$).

The following result summarizes the more salient experimental findings about behavior and beliefs under complete *vs.* incomplete information.

Result 7 Polarization of subjects' behavior and beliefs due to questionnaire disclosure in phase 3 of *QD* is observed both by taking phase 1 of *QD* and by taking phase 3 of *NoQ-QnoD* as controls. The most significant difference is found for *A*-subjects matched with high-guilt *B*-subjects in phase 3 of *QD*.

5 Concluding Remarks

The main innovation of the experimental design of this paper (and of the twin project of Attanasi *et al.* 2019b) is to make *B*-subjects in a Trust Minigame (the trustees) answer a structured questionnaire that reveals their psychological type. In the main treatment we make the

filled-in questionnaire common knowledge within the matched pair. Under the assumption that A, the truster, is commonly known to be selfish, we interpret the treatment-*vs*-control comparison as one between complete and incomplete information. A second innovation of this paper is that we organize the data with an original theoretical analysis integrating guilt aversion and reciprocity, and we obtain both rationalizability and equilibrium predictions under the two information regimes. The predictions of rationalizability are coarse and yield differences in behavior according to the information regime only for A subjects. The equilibrium predictions refine the rationalizability predictions and yield differences in behavior also for intermediate psychological types of B that do not belong to forward-induction dominance regions. Such differences can only occur under belief-dependent preferences, because a B-subject who only cares about the material payoffs allocation (e.g., an inequity averse type) has a weakly dominant strategy, which is independent of the information regime. Yet, since only the intermediate types of B change behavior across information regimes, we predict smaller changes in choice distributions for B-subjects than for A-subjects.

Our theoretical analysis of B's answers to the questionnaire (Proposition 1) is able to capture the great majority of B-subjects' payback patterns, and most of them are belief-dependent (Result 1). Among these belief-dependent types, we find that all B-subjects predicted to share in the complete-information equilibrium (Proposition 3) are "high-guilt" types, i.e., types for whom the other-regarding attitude $G + R$ is above the theoretical threshold, and the guilt component G is large in absolute terms and relative to the reciprocity component R (Result 2).

Our theoretical predictions capture well the central tendencies of the data (Result 3 for the complete-information treatment-phase combination, and Result 6 for the other—incomplete-information— combinations): The complete-information theory summarized in Propositions 2 and 3 implies the polarization of behavior and beliefs found only in phase 3 of the QD (questionnaire disclosure) treatment, where the psychological type estimated from the questionnaire filled in by B correlates in the predicted direction with the choices and beliefs of A (Result 7). By contrast, A's and B's choices are statistically independent in the other phase-treatment combinations, as expected in a stranger matching setting.

More precisely, high-guilt estimated types of B are more likely to share than low-guilt ones (Results 5 and 6), and A-subjects are more likely to trust under questionnaire disclosure when matched with a high-guilt B-subject rather than with a low-guilt one (Result 4). Furthermore, B's propensity to share is higher under questionnaire disclosure (Result 7), as predicted by our equilibrium model of belief-dependent preferences (Propositions 3 and 5).

The model predicts very well A's propensity to trust when she is matched with a high-guilt B-subject in phase 3 of QD (Result 7). The most important deviation from the model is that B's high-guilt types share much less than predicted (Results 3 and 6). Our informed

conjecture is that this deviation is in part due to lower than predicted conditional second-order beliefs, which we cannot measure accurately. Attanasi *et al.* (2016) provide a possible explanation: If A, like B, is potentially guilt averse, B may interpret action *Continue* as a desire of A not to disappoint B; hence, it may well be the case that β, B's estimate of A's belief in *Share*, is less than 1/2. Indeed, other experimental works on the Trust Minigame (e.g., Charness & Dufwenberg 2006) show that a significant fraction of B-subjects hold such low conditional second-order beliefs.

References

Attanasi, G., and R. Nagel (2008): "A Survey of Psychological Games: Theoretical Findings and Experimental Evidence," in *Games, Rationality and Behavior. Essays on Behavioral Game Theory and Experiments*, A. Innocenti and P. Sbriglia (Eds.). Houndmills: Palgrave McMillan, 204–232.

Attanasi, G., P. Battigalli, and E. Manzoni (2016): "Incomplete Information Models of Guilt Aversion in the Trust Game," *Management Science*, 62, 648–667.

Attanasi, G., C. Rimbaud, and M. C. Villeval (2019a): "Embezzlement and Guilt Aversion," *Journal of Economic Behavior & Organization*, 167, 409–429.

Attanasi, G., C. Rimbaud, and M. C. Villeval (2020): "Guilt Aversion in (New) Games: the Role of Vulnerability," *GREDEG Working Papers 2020-15*, Université Côte d'Azur, France.

Attanasi, G., C. Rimbaud, and M. C. Villeval (2022): "Guilt Aversion in (New) Games: Does Partners' Vulnerability Matter?," *Available at SSRN*: https://ssrn.com/abstract=4066820

Attanasi, G., P. Battigalli, E. Manzoni, and R. Nagel (2019b): "Belief-dependent Preferences and Reputation: Experimental Analysis of a Repeated Trust Game," *Journal of Economic Behavior & Organization*, 167, 341–360.

Bacharach, M., G. Guerra, and D. J. Zizzo (2007): "The Self-Fulfilling Property of Trust: An Experimental Study," *Theory and Decision*, 63, 349–388.

Balafoutas, L., and H. Fornwager (2017): "The Limits of Guilt," *Journal of the Economic Science Association*, 3, 137–148.

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

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

Battigalli, P., and M. Dufwenberg (2022): "Belief-Dependent Motivations and Psychological Game Theory," *Journal of Economic Literature*, in print.

Battigalli, P., and M. Siniscalchi (2002): "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory,* 106, 356–391.

Battigalli, P., R. Corrao, and M. Dufwenberg (2019a): "Incorporating belief-dependent motivation in games," *Journal of Economic Behavior & Organization*, 167, 185–

218.

Battigalli, P., R. Corrao, and F. Sanna (2020): "Epistemic Game Theory Without Type Structures. An Application to Psychological Games," T*Games and Economic Behavior*, 120, 28–57.

Battigalli, P., M. Dufwenberg, and A. Smith (2019b): "Frustration, Aggression, and Anger in Leader-follower Games," *Games and Economic Behavior*, 117, 15–39.

Bellemare, C., A. Sebald, and M. Strobel (2011): "Measuring the Willingness to Pay to Avoid Guilt: Estimation Using Equilibrium and Stated Belief Models," *Journal of Applied Econometrics*, 26, 437–453.

Bellemare, C., A. Sebald, and S. Suetens (2017): "A Note on Testing Guilt Aversion," *Games and Economic Behavior*, 102, 233–239.

Bellemare, C., A. Sebald, and S. Suetens (2018): "Heterogeneous Guilt Aversion and Incentive Effects," *Experimental Economics*, 21, 316–336.

Berg, J., J. Dickhaut, and K. McCabe (1995): "Trust, Reciprocity, and Social-History," *Games and Economic Behavior*, 10, 122–142.

Berkowitz, L., and E. Harmon-Jones (2004): "Toward an Understanding of the Determinants of Anger," *Emotion*, 4, 107–130.

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

Brañas-Garza P., M. Bucheli, M. P. Espinosa, and T. Garcia-Muñoz (2013): "Moral Cleansing and Moral Licenses: Experimental Evidence," *Economics and Philosophy*, 29, 199–212.

Buskens, V., and W. Raub (2013): "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust," in *Handbook of Rational Choice Social Research*, ed. by R. Wittek, T. A. B. Snijders, and V. Nee. New York: Russell Sage, 113–150.

Cartwright, E. (2019): "A Survey of Belief-based Guilt Aversion in Trust and Dictator Games," *Journal of Economic Behavior and Organization*, 167, 430–444.

Chang, L. J., A. Smith, M. Dufwenberg, and A. Sanfey (2011): "Triangulating the Neural, Psychological and Economic Bases of Guilt Aversion," *Neuron*, 70, 560–572.

Chao, M. (2018): "Intentions-based Reciprocity to Monetary and Non-monetary Gifts," *Games*, 9(4), 74.

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

Charness, G., and M. Dufwenberg (2011): "Participation," *American Economic Review*, 101, 1213–1239.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Charness, G., A. Samek, and J. van de Ven (2022): "Understanding Social Preferences with Simple Tests," *Experimental Economics*, 25, 385–412.

Costa-Gomes, M., V. P. Crawford, and B. Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 1193–1235.

Danilov, A., K. Khalmetski, and D. Sliwka (2021): "Descriptive Norms and Guilt Aversion," *Journal of Economic Behavior & Organization*, 191, 293–311.

Dekel E., and M. Siniscalchi (2015): "Epistemic Game Theory," in P. Young and S. Zamir (Eds.), *Handbook of Game Theory*, 4, 619-702. Amsterdam: North Holland (Elsevier).

Dhaene, G., and J. Bouckaert (2010): "Sequential Reciprocity in Two-player, Two-stage Games: An Experimental Analysis," *Games and Economic Behavior*, 70, 289–303.

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

Dufwenberg, M. (2002): "Marital Investment, Time Consistency and Emotions," *Journal of Economic Behavior & Organization*, 48, 57–69.

Dufwenberg, M. (2008): "Psychological Games," in *The New Palgrave Dictionary of Economics,* 6, ed. by S. N. Durlauf and L. E. Blume, 714–718.

Dufwenberg, M., and U. Gneezy (2000): "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior*, 30, 163–182.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

Dufwenberg, M., S. Gachter, and H. Hennig-Schmidt (2011): "The Framing of Games and the Psychology of Play," *Games and Economic Behavior*, 73, 459–478.

Dufwenberg, M., A. Smith, and M. Van Essen (2013): "Hold-up: with a vengeance," *Economic Inquiry*, 51, 896–908.

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

Elster, J. (1998): "Emotions and Economic Theory," *Journal of Economic Literature*, 36, 47–74.

Engler, Y., R. Kerschbamer, and L. Page (2018): "Guilt averse or reciprocal? Looking at behavioral motivations in the trust game," *Journal of the Economic Science Association*, 4, 1–14.

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

Falk, A., E. Fehr, and U. Fischbacher (2008): "Testing Theories of Fairness - Intentions Matter," *Games and Economic Behavior*, 62, 287–303.

Fischbacher, U. (2007): "Z-Tree: Zurich Toolbox for Readymade Economic Experiments," *Experimental Economics*, 10, 171–178.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60–79.

Gómez-Miñambres, J., E. Schniter, and T. W. Shields (2021): "Investment Choice Architecture In Trust Games: When "All-In" Is Not Enough," *Economic Inquiry*, 59, 300–314.

Guerra, G., and D. J. Zizzo (2004): "Trust Responsiveness and Beliefs," *Journal of Economic Behavior and Organization*, 55, 25–30.

Harsanyi J. (1967-68): "Games of Incomplete Information Played by Bayesian Players. Parts I, II, III," *Management Science*, 14, 159–182, 320–334, 486–502.

Healy, P. (2011): "Epistemic Foundations for the Failure of Nash Equilibrium," Typescript, Ohio State University.

Jensen, M. K. and M. Kozlovskaya (2016): "A representation theorem for guilt aversion," *Journal of Economic Behavior Organization*, 125, 148–161.

Ketelaar, T., and W. T. Au (2003): "The Effects of Feelings of Guilt on the Behavior of Uncooperative Individuals in Repeated Social Bargaining Games: An Affect-as-Information Interpretation of the Role of Emotion in Social Interaction," *Cognition and Emotion*, 17, 429–453.

Khalmetski, K. (2016): "Testing guilt aversion with an exogenous shift in beliefs," *Games and Economic Behavior*, 97, 110–119.

Khalmetski, K., A. Ockenfels, and P. Werner (2015): "Surprising Gifts: Theory and Laboratory Evidence," *Journal of Economic Theory*, 159, 163–208.

Morell, A. (2019): "The Short Arm of Guilt – An Experiment on Group Identity and Guilt Aversion," *Journal of Economic Behavior & Organization*, 166, 332–345.

Orhun, A. Y. (2018): "Perceived Motives and Reciprocity," *Games and Economic Behavior*, 109, 436–451.

Peeters, R., and M. Vorsatz (2021): "Simple guilt and cooperation," *Journal of Economic Psychology*, 82, 102347.

Podsakoff, P. M., S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff (2003): "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology*, 88, 879–903.

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

Regner, T., and N. S. Harth (2014): "Testing Belief-dependent Models," Working Paper, Max Planck Institute of Economics, Jena.

Reuben, E., P. Sapienza, and L. Zingales (2009): "Is Mistrust Self-Fulfilling?" *Economic Letters*, 104, 89–91.

Rimbaud, C., and A. Soldà (2021): "Avoiding the Cost of your Conscience: Belief Dependent Preferences and Information Acquisition," Working Paper 2114, GATE, Lyon.

Sachdeva, S., R. Iliev, and D. Medin (2009): "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation," *Psychological Science*, 20, 523–528.

Schotter, A., and I. Trevino (2014): "Belief Elicitation in the Lab," *Annual Review of Economics*, 6, 103–128.

Silfver, M. (2007): "Coping with Guilt and Shame: A Narrative Approach," *Journal of Moral Education*, 36, 169–183.

Siniscalchi, M. (2016): "Structural Rationality in Dynamic Games," Typescript, Northwestern University.

Smith, A. (1759): *The Theory of Moral Sentiments*. London: A. Millar.

Stanca, L., L. Bruni, and L. Corazzini (2009): "Testing Theories of Reciprocity:

Do Motivations Matter?" *Journal of Economic Behavior & Organization*, 71, 233–245.

TOUSSAERT, S. (2017): "Intention-Based Reciprocity and Signalling of Intentions," *Journal of Economic Behavior & Organization,* 137, 132–144.