Disclosure of Belief–Dependent Preferences in a Trust Game[*]

Giuseppe Attanasi (BETA, University of Strasbourg)

Pierpaolo Battigalli (Bocconi University and IGIER, Milan)

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

February 2016

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. 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. 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, incomplete information.

[*] We thank Olivier Armantier for great support in the statistical analysis. We thank Stefania Bortolotti, Martin Dufwenberg, Alejandro Francetich, Pierfrancesco Guarino, Elena Manzoni, Sara Negrelli, Salvatore Nunnari, Fabrizio Panebianco, Jacopo Perego, Ariel Rubinstein, Fabio Tufano, and the seminar participants at CISEI 2010 in Capri, ESA 2010 in Copenhagen, BQGT 2010 in Newport, IMEBE 2011 in Barcelona, 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 for useful discussions. 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 2009, 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).[1]

We study belief-dependent motivations in the Trust Game, a stylized social dilemma whereby agent A (the truster) takes a costly action that generates a social return, and agent B (the trustee) 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 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.[2]

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). According to **intention-based reciprocity**, instead, 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 costly action; therefore, B's willingness to share is *decreasing* in his second-order belief (Dufwenberg & Kirchsteiger 2004).[3] Experimental studies of the Trust Game find a positive correlation

[1] There are other important sources of belief-dependent motivation. In particular, agents may have a non-instrumental concern for esteem (e.g. Ellingsen & Johanneson 2008) and self-esteem (e.g. Kuhnen & Tymula 2012), which are determined by beliefs. Tadelis (2011) experimentally studies the effect of exposure in the Trust Game. His results are consistent with a theory of shame avoidance.

[2] From now on we refer to A (B) as a female (male).

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

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 & Dufwenberg 2006, Chang *et al.* 2011). However, other experimental studies find evidence in support of intention-based reciprocity both in the Trust Game (Cox 2004, Bacharach *et al.* 2007, Stanca *et al.* 2009, Toussaert 2015) and in other social dilemmas (e.g., Falk *et al.* 2008). Thus, the experimental evidence suggests that both motivations are present in human interaction, although it seems that guilt aversion prevails for a majority of the non-selfish B-subjects in the Trust Game.[4]

A common feature 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.[5] Theoretical predictions are harder to derive for such games, because incomplete-information models are more complex, and their specification is more arbitrary. Indeed, on top of the distribution of preferences, the analyst also has to specify the distribution of possible hierarchical beliefs about such preferences, that is, the beliefs about the co-player preferences, beliefs about such beliefs, and so on (see Harsanyi 1967-68). Theoretical analysis and introspection give some guidance on the specification of preferences, but little guidance on the specification of belief hierarchies.[6]

To see the relevance of (in)completeness of information, assume for simplicity that preferences 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 or her matched pair. In this hypothetical complete-information regime, information about B would work as a correlating device selecting either the cooperative outcome (when B is known to be trustworthy), or the no-trust outcome (when B is known to be untrustworthy, A does not take the costly, surplus increasing action). In particular, we would rarely observe B grabbing the surplus created by A's

games. See also Charness & Rabin (2002), Falk & Fischbacher (2006), and Stanca *et al.* (2009).

[4] See Guerra & Zizzo (2004), Bacharach *et al.* (2007), Charness & Dufwenberg (2011), Bracht & Regner (2013) and Ederer & Stremitzer (2015) for the Trust Game. Dufwenberg & Gneezy (2000), Reuben *et al.* (2009) and Bellemare *et al.* (2011) find support for guilt aversion in other social dilemmas. Vanberg (2008), Ellingsen *et al.* (2010) and Kawagoe & Narita (2014) instead, do not find evidence corroborating the guilt-aversion hypothesis.

[5] 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**.

[6] For an incomplete information analysis of the Trust Minigame with guilt aversion and some more general considerations about psychological games with incomplete information, see Attanasi *et al.* (2016).

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's subjects are optimistic enough to trust B, and 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 complete and incomplete information regimes. If, instead, B-subjects have belief-dependent preferences (like guilt-aversion or intention-based reciprocity), then we should expect 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.

Our study addresses these issues theoretically and experimentally. We do make the above-mentioned simplifying assumption that the truster, A, is self-interested; on the other hand, the trustee, B, has belief-dependent preferences given by a combination of guilt aversion and intention-based reciprocity. We elicit the trustee's belief-dependent preferences through a structured questionnaire.[7] 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, this treatment implements a psychological game with complete information.

[7] Bellemare *et al.* (2015) and Khalmetski *et al.* (2015) elicit the dictator's belief-dependent preferences in a dictator game through a structured questionnaire comparable 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 derive precise predictions for the complete-information model, and robust qualitative predictions for the incomplete-information model. Roughly, if the trustee is highly guilt averse, or reciprocal, the Pareto-dominating equilibrium is cooperative, with high first- and second-order beliefs, whereas in the opposite case the unique equilibrium is un-cooperative, with low first- and second-order beliefs. Thus, disclosing B's type acts as a correlation device. Under incomplete information, instead, random matching decouples A's and B's behavior/beliefs. As for the latter, the threshold for guilt aversion above which B shares is higher than under complete information. Thus, we predict a different distribution not only of A-strategies, but also of B-strategies (and beliefs) in the two treatments. In particular, if guilt-aversion is the prevalent motivation (something we test), we predict more trust and more sharing under disclosure of the belief-dependent preferences.

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, the trustee's propensity to share is indeed increasing with 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.

The rest of the paper is structured as follows. In Section 2 we describe our experimental design. Section 3 presents our theoretical model. In Section 4 we present and discuss our experimental results in light of the theoretical predictions. An Online Appendix collects technical details about the experimental instructions, the theoretical model and raw experimental data.[8]

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 on the design (2.3).

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 that has thus far yielded a total profit 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 contract states that

[8] See, respectively, *Online Appendix A*, B and C of the paper, at http://www.igier.unibocconi.it/wp506.

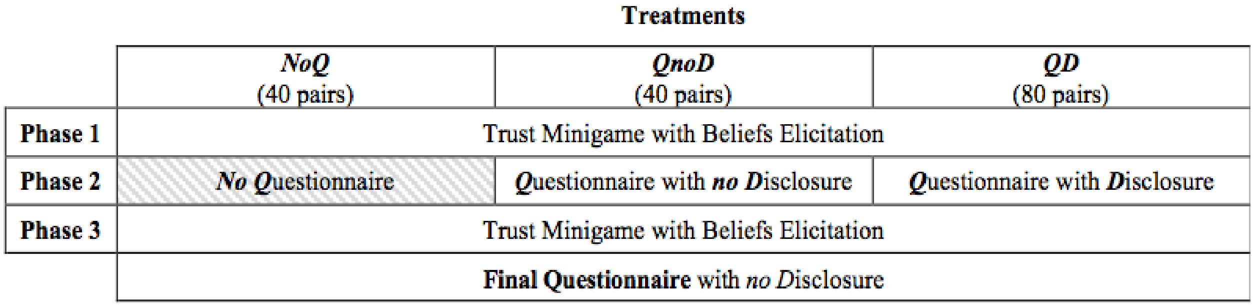
the players split the profit fifty-fifty. If player A decides to *Continue* with the partnership, total profit doubles (€4); however, in that case, player B has the right to share equally or take the whole surplus. In the simultaneous-move game of Table 1 (the strategic form of the Trust Minigame), player B – before knowing player A's choice – has to state if he would *Take* or (equally) *Share* the higher profits.

	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 1st and 2nd year undergraduate students in Economics at Università Bocconi of Milan. 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.



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 D*isclosure		

Table 2 Summary of the Experimental Design.

Design The experimental design is made of three phases and three treatments, explained in detail in Table 2.[9] The difference between treatments is in phase 2, depending on whether subjects playing in the role of player B are asked to fill a questionnaire and whether such

[9] The English translation of the instructions is provided as an electronic supplementary material of this paper: see *Online Appendix A* of the paper, at http://www.igier.unibocconi.it/wp506.

answers are disclosed to subjects playing in the role of player A. 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 the experiment, each participant is randomly chosen to play in role A or role B of the Trust Minigame. He/she maintains the same role until the end of the experiment. Subjects in role A (henceforth A-subjects) are randomly matched with subjects in role B (henceforth B-subjects), thus creating 10 matched pairs in each session.

Participants are told that the experiment is made of three phases. Instructions of the new phase are given and read aloud only prior to that phase. Phase 1 consists of two decision tasks: the Trust Minigame of Table 1 and a belief-elicitation task described in detail below. Phase 2 consists of a new random matching and, in *QnoD* and *QD,* also of the *Questionnaire* explained below. Phase 3 consists of the same decision tasks as phase 1, but with a new random matching (stranger matching design). In particular, in *QnoD* and *QD,* the random matching is the same as in phase 2.

After phase 3, there is another questionnaire, the same one for all treatments and equal to the one in phase 2 of *QnoD* and *QD.* Only after this *Final Questionnaire* are subjects told the results of phase 1 and phase 3, and are paid the sum of the earnings in these two phases. We now describe in detail the three phases of the experimental design.

Phase 1 This phase is the same for all treatments, and consists of two decision tasks. First, elicitation of beliefs about strategies and beliefs in 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).[10] Each B-subject is asked to *guess the answer* of his co-paired A about the percentage of B-subjects who will choose *Share* (B's unconditional second-order belief),[11] and to *guess the choice* that his co-paired A will make (a feature of B's first-order belief).

Then, within each pair, player A and player B simultaneously make their choice in the strategic form of the Trust Minigame of Table 1. Subjects do not receive any information feedback at the end of phase 1. Indeed, at the beginning of the phase, subjects are informed that the resulting payoffs of the Trust Minigame and the gains for the correctness of the beliefs will both be communicated at the end of the experiment.

[10] We ask A to indicate a percentage $\pi\%$, with $\pi = 10 \cdot n$ and $n \in \{0, 1, 2, ..., 10\}$, since it is public information that there are 10 B-subjects in each session.

[11] As we did for player $A,$ we ask player B to indicate a percentage $\pi\%$, with $\pi = 10 \cdot n$ and $n \in \{0, 1, 2, ..., 10\}$.

Phase 2 In *NoQ* subjects proceed directly to phase 3. In *QnoD* and *QD*, subjects are randomly re-matched to form other 10 pairs. B-subjects are asked to fill the questionnaire of Table 3. In particular, each B 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 the €4 back to A. He is allowed to condition his payback on the new A-co-player's guess of percentage of B-subjects choosing *Share*.

Since there are 10 B-subjects, when choosing *Continue*, A had 11 possible guesses about how many B-subjects could choose *Share* (0%, 10%, ..., 100%), as shown in Table 3.[12] Hence, each B-subject is asked to insert a value (between €0.00 and €4.00) in each of the 11 rows of Table 3.

A's possible assessments of *Share*	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.

The following features of phase 2 are made public information among the subjects in a session of *QnoD* and *QD*: Neither the responding subject nor anyone else will receive any payment for the answers he/she gives in the questionnaire of Table 3 (hypothetical payback scheme); A-subjects read and listen to the instructions of phase 2; B-subjects fill the questionnaire first on a sheet of paper and then copy the answers on the questionnaire in their computer screen. Furthermore, in *QnoD* it is public information that B's filled-in questionnaire *will not be* disclosed to anyone, while in *QD* it is public information that B's filled-in questionnaire *will be* disclosed to a randomly-chosen A-subject. In fact, in *QD*, player A receives the filled-in

[12] To check for framing effects, in half of the sessions of each treatment, the first column of Table 3 is shown in reverse order, with 100% on the first row and 0% on the last row.

questionnaire of her new co-player B, randomly chosen at the beginning of phase 2. More precisely, at the end of phase 2, B's filled-in questionnaire appears on A's screen, and the latter is invited to copy it on a sheet of paper.

Phase 3 The two decision tasks are the same as in phase 1. 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. Furthermore, each B-subject can keep his previously filled-in paper questionnaire in front of him in this phase. Additionally, in this phase of *QD*, A can keep the paired B's filled-in questionnaire (previously copied on a sheet of paper) in front of her. At the beginning of phase 3, 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.

As in phase 1, in all treatments subjects do not receive any information feedback at the end of phase 3.

Final questionnaire After phase 3, there is a final questionnaire, which is the same for all treatments (see Table 3): In *QnoD* and *QD*, we ask B-subjects to fill the same questionnaire on a sheet of paper as in phase 2, knowing that it *will not be* disclosed to anyone.

In *NoQ*, this is the first time B-subjects fill the questionnaire of Table 3; in *QnoD* and in *QD* – being the second time B-subjects face the same questionnaire – they are allowed to give answers different from those given 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 (player A) or his first- and second-order beliefs (player B) in phase 1 and in phase 3 were correct.

Each subject is paid the sum of the resulting payoffs in the Trust Minigame in phase 1 and in phase 3, and is also paid for correct guesses (elicited beliefs). Specifically, €5 are added to the total payoff of A-subjects for each correct first-order belief (in phases 1 and 3). Similarly, €5 are added to the total payoff of each B-subject for every time he guessed correctly both the choice and the first-order belief of the co-player (in phases 1 and 3).

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 The one-shot interaction that is relevant for the comparison between treatments is phase 3, i.e. after B-subjects are asked (or not) to fill the questionnaire of Table 3 and such answers are disclosed (or not) to A-subjects. However, 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.

Second, we want to let subjects understand the Trust Minigame and the belief-elicitation procedure before B-subjects fill the questionnaire in phase 2 of *QnoD* and *QD*. Indeed, in each session there are 10 B-subjects, hence 11 possible values for the frequency of *Share* choices in phase 1: Each of these values corresponds to one of the 11 rows of the questionnaire, which makes it more salient.

For *NoQ*, phase 1 has been mainly introduced to maintain the same structure as in *QnoD* and *QD*, thereby making players' behavior in phase 3 comparable between *NoQ* and the other 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.[13]

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 a simultaneous game: the strategic form of the Trust Minigame.[14] Differently from them, and similarly to Guerra & Zizzo (2004), we elicit beliefs before choices. Eliciting beliefs first should not change behavior in the subsequent Trust Minigame. Indeed, Guerra & Zizzo (2004) find no difference in trust and cooperation between comparable treatments with and without beliefs elicitation when subjects play a Trust Minigame similar to ours with the strategy method.

First-order beliefs of A-subjects are elicited as in Charness & Dufwenberg (2006) and follow-up papers on the Trust Minigame.[15] Like them, we do not ask A to guess the likelihood that the paired B would choose *Share*, since we do not observe this likelihood: The observed binary choice would make this simply a *Yes* or *No* guess. Instead, we ask A to guess how

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

[14] 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 2014). But we think that our description of the game in the experimental instructions essentially avoids such framing effects (see *Online Appendix A*).

[15] See, e.g., Bracht & Regner (2013).

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*: They ask B to state A's average guess of the percentage of B-subjects choosing *Share*, by considering only those A-subjects choosing *Continue*. The main reason for eliciting 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 the number of rows in the questionnaire to match the number of B-subjects in the session. Thus, in order to have manageable number of rows in the questionnaire, we only have 10 B-subjects in each session. This is too small a number for making a reliable inference about what A-subjects think, if one considers only A-subjects who choose *Continue*.

We now offer some theoretical considerations about the relevance of conditional and unconditional beliefs. Our subjects play a simultaneous game, but the choice of B is equivalent to a contingent plan in the sequential version of the game, and – therefore – should correlate with his conditional belief. However, unconditional beliefs are relevant as well, because they reflect how players reason strategically before playing the game.[16] Indeed, our incomplete-information model of subsection 3.3 provides some testable predictions about unconditional beliefs in treatments *NoQ* and *QnoD* (see Proposition 3). This also motivates our elicitation of B's unconditional *first*-order beliefs, unlike most previous experimental studies on the Trust Minigame.[17] 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.[18] 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 beliefs and the actions of the co-player.

Information in Phase 2 In both treatments *QnoD* and *QD*, A-subjects read and listen to the instructions of phase 2: This is made for A-subjects to know what B-subjects are asked to do in phase 2, and in *QD* also to help them interpret the disclosed filled-in questionnaire. The reason for asking subjects in role B (A) to fill in (copy) the questionnaire on a sheet

[16] 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.

[17] As 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.

[18] 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 $\mathbb{P}_B(\alpha \leq x | Cont.) = \mathbb{P}_B(\alpha \leq x)$.

of paper in phase 2 is twofold. First, by asking B-subjects in *QnoD* and *QD* to fill in the questionnaire twice, we make them think more carefully about their answers, and control for the correspondence between answers in the paper and in the computerized questionnaire. Similar considerations apply to A-subjects in *QD*: Asking them to copy B's answers on a sheet of paper should make them focus on these answers.

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 in the questionnaire will be used later. Specifically, in phase 2 it is public information that the filled-in form *will be disclosed* to a randomly-chosen A-subject, but only at the beginning of phase 3 is it made public information within each pair that the randomly-chosen 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 questionnaire recipient.

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 filled in the questionnaire in phase 2 change or confirm their answers (in *QnoD* and *QD*).[19] In the latter case, we cannot reject the hypothesis that subjects truthfully revealed their belief-dependent preferences.

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 (3.2), and predictions for the Trust Minigame (3.3), both under complete information (3.3.1) and incomplete information (3.3.2).

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

[19] 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 in phase 2 when filling in the final questionnaire. Leaving this paper with them could have biased the answers to the final questionnaire.

(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. Assuming that j has a deterministic plan, viz. strategy s_j, α_j is determined by the pair (s_j, α_{ji}), where α_{ji} is j's 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 $\mathbb{E}_A[\widetilde{m}_A; \alpha_A] = 2 \times 0.6 = 1.2$. 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 $u_B(m_B, m_A, \widetilde{\alpha}_A)$.

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 reciprocity concerns: If j is kind (unkind), i is willing to sacrifice some monetary payoff to increase (decrease) the monetary payoff of j.

In the "simple-guilt" model of Battigalli & Dufwenberg (2007), utility depends linearly on both monetary payoff and disappointment, and the reciprocity term is zero. This model has been mostly used to analyze binary allocation choices, such as the choice of player B (the trustee) in the Trust Minigame. Here, instead, we rely on belief-dependent preferences also to analyze the payback scheme shown in Table 3 above, where B-subjects answer hypothetical questions by choosing distributions in a fine grid that approximates a continuum. With this goal in mind, in both choices that B is asked to make in our experimental setting – the Trust Minigame and the hypothetical payback scheme –, we use a parametric specification of (1) where 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, and the reciprocity term $r_i(\cdot)$ is linear:

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

where G_i and R_i respectively parametrize sensitivity to guilt and reciprocity. This parametrization achieves a good balance between tractability and flexibility.

In the context of the Trust Minigame, we assume that A (the truster) has *selfish* risk-neutral preferences. 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, α_{AB}. To simplify notation, from now on we let $\alpha = \mathbb{P}_A(Share)$ denote this variable, and $\beta = \mathbb{E}_B(\alpha|Cont.)$ 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 in Battigalli and 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

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

However, there is a possible confound. Since we put the B responder in a hypothetical situation in which he has "transgressed," 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).[20] Therefore, 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

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

By strict concavity, (4) has a unique solution $x^* = \xi(\alpha)$. We call $\xi(\alpha)$ the **payback function.**[21] The first-order condition for an interior solution helps us understand how the payback changes as a function of the first-order belief α and of parameter shifts. It is useful to think in terms of the "marginal cost" and "marginal benefit" of the payback x:

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

Next we describe the main features of the payback function $\xi(\alpha)$ and its dependence on

[20] 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.

[21] The *Online Appendix B* of the paper, at http://www.igier.unibocconi.it/wp506, contains a derivation of the payback function $\xi(\alpha)$ in closed form; here we provide intuition.

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) constant *if $G=R$ and $R\leq \underline{R}\,(p)$,*
*(iii) first decreasing and then increasing (*U-shaped*) if $G>R$ and $\underline{R}\,(p)<R<\overline{R}\,(p)$,*
(iv) decreasing *if either $G<R$ or $R\geq \overline{R}\,(p)$,*
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 $\xi(\alpha)<2p+2\alpha$.

These results can be understood drawing the MC and MB schedules under different cases and tracing how their intersection is affected by parameter shifts.[22] Proposition 1 describes the four possible shapes of the payback function. Roughly, the function is increasing when guilt aversion prevails on reciprocity, it is constant when they are balanced, it is decreasing when reciprocity prevails on guilt aversion, and it is U-shaped when $G>R$ and R has intermediate values. The intuition for the latter 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. More formally, Proposition 1 implies that $\xi(\alpha)$ is locally increasing (hence it is an interior solution) iff $G>R$ and $0<\xi(\alpha)<2p+2\alpha$, which follows from the implicit function theorem: An interior solution $x^*=\xi(\alpha)\in(0,4)$ satisfies the first-order condition (5); differentiating it, we get[23]

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

3.3 Equilibrium analysis of the Trust Minigame

Since we assume that B-subjects choose as if they had observed the trusting action *Continue*, we analyze the perfect Bayesian equilibria (PBE) of the sequential Trust Minigame, a game

[22] See *Online Appendix B* of the paper, at http://www.igier.unibocconi.it/wp506.

[23] 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 sufficiently high "simple guilt."

with perfect information.[24] We consider two situations: the complete-information benchmark of common knowledge of the psychological utility function u_B in (2), which we claim we approximate in the lab in the main treatment, and the incomplete-information case where u_B is not common knowledge, which is the standard situation in experiments.

It is well-known that psychological games have multiple PBEs 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 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.[25]

3.3.1 Complete information

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.[26] 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 $\beta := \mathbb{E}_B[\widetilde{\alpha}|Cont.] = \alpha$.

The simultaneous presence of the guilt and reciprocity components in model (2) implies that the propensity to share can either increase or decrease with the conditional second-order belief β. In particular, if R is sufficiently high, B may prefer *Take* when $\beta = 1$ and *Share* for lower values of β. As we show below, this creates the possibility of a partially randomized equilibrium where A chooses *Continue*, B is indifferent, and he chooses *Share* with probability $\alpha \geq 1/2$ (cf. Battigalli & Dufwenberg 2009, subsection 4.3.3).

Plugging the disappointment and kindness functions in (2), we obtain

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

[24] 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.

[25] This is also the equilibrium with higher psychological utility. Notice that forward-induction reasoning selects the same equilibrium when the guilt component is high enough (Dufwenberg 2002, Attanasi & Nagel 2008, Battigalli & Dufwenberg 2009). However, there is a range of intermediate values for which the "trusting equilibrium" exists, but it is not selected by forward induction.

[26] 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 $\mathbb{P}_B[\widetilde{\alpha} = \alpha|Cont.] = 1 = \mathbb{P}_B[\widetilde{\alpha} = \alpha]$.

where $(m_A, m_B) = (2,2)$ if player B chooses *Share* and $(m_A, m_B) = (0,4)$ if he chooses *Take*. Therefore, player B chooses *Share* iff $u_B(2,2,\alpha) \geq u_B(4,0,\alpha)$ according to (6), that is,

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

As explained above, in a complete-information equilibrium where A chooses *Continue*, the conditional second-order belief of B assigns probability one to the true value α. Therefore, inequality (7) becomes

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

Our equilibrium analysis depends on the shape of the **"willingness-to-share" function** $WS(\alpha; G, R)$ implied by the (psychological) utility type (G, R) (see Figure 1).

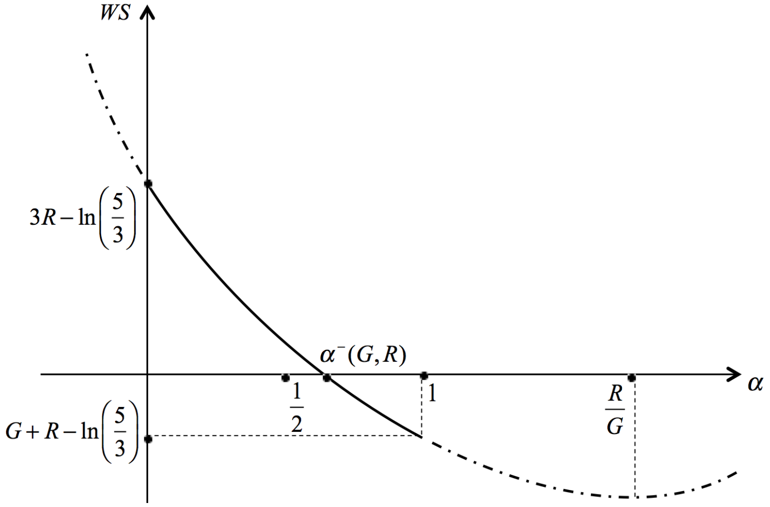


Figure 1 B's equilibrium willingness to share for $G{+}R$ small and R/G large.

In particular, $(Continue, Share, \alpha = \beta = 1)$ is an equilibrium – hence the Pareto-superior equilibrium – iff $G + R \geq \ln(5/3) \approx 0.52$. More generally, there is an equilibrium where A chooses *Continue* with positive probability iff (7) holds when B assigns probability one to the true value of α ex ante, and conditional on *Continue*, and α must be larger than 1/2 (otherwise A would choose *Dissolve*). This holds iff $WS(\alpha; G, R) \geq 0$ for $\alpha \geq 1/2$.

The quadratic, convex function $WS(\alpha; G, R)$ (with unrestricted $\alpha \in \mathbb{R}$) has a minimum at $\alpha = R/G$, and attains value $WS(1; G, R) = G{+}R{-}\ln(5/3)$ at $\alpha = 1$. If (*Continue*,*Share*) is not an equilibrium, that is, if $WS(1; G, R) < 0$, the equation $WS(\alpha; G, R) = 0$ necessarily has two real-valued solutions, the largest of which must be larger than one. Therefore, only the smaller solution,

$$\alpha^{-}(G,R)=\frac{R-\sqrt{R^{2}-G\left(3R-\ln\left(5/3\right)\right)}}{G}\ , \tag{9}$$

matters for the analysis. At $\alpha=\alpha^{-}(G,R)$, the graph of $WS(\alpha;G,R)$ cuts the horizontal axis from above.

If $\alpha^{-}(G,R)\geq 1/2$, then there is an equilibrium where A chooses *Continue* and correctly believes that B – who is indifferent – chooses *Share* with probability $\alpha=\alpha^{-}(G,R)$. This is also the Pareto-superior equilibrium.[27] If $\alpha^{-}(G,R)<1/2$, A chooses *Dissolve* in equilibrium. If $\alpha^{-}(G,R)\leq 0$, the equilibrium strategy of B is *Take*. If $0<\alpha^{-}(G,R)<1/2$, there are multiple, payoff-equivalent equilibria where the probability of *Share* (which is equal to the first-order belief α) is lower than $1/2$.[28]

Replacing the expression for $\alpha^{-}(G,R)$ of eq. (9) in the above inequalities, we obtain $\alpha^{-}(G,R)\geq 1/2$ iff $R\geq\ln\left(5/3\right)/2-G/8$, and $\alpha^{-}(G,R)\leq 0$ iff $R\leq\ln\left(5/3\right)/3$. Therefore, the Pareto-superior equilibrium depends on G and R as follows (see Figure 2):

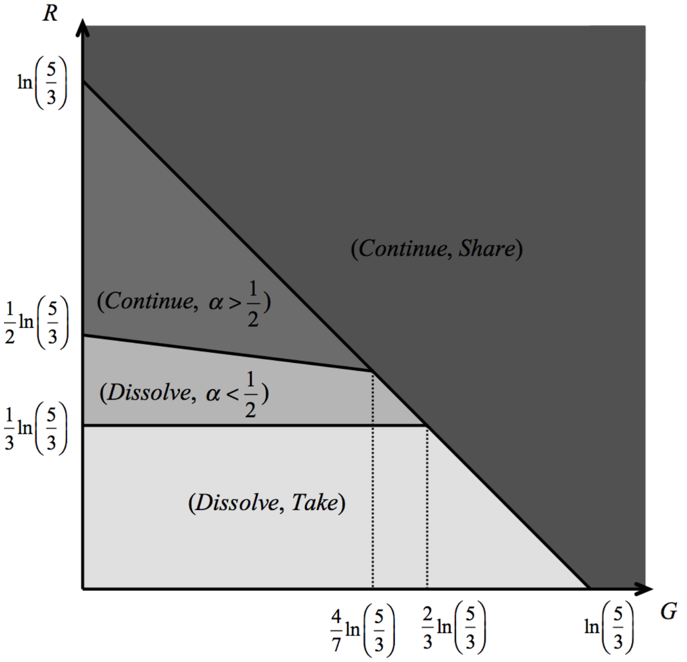

R
$\ln\left(\frac{5}{3}\right)$
$\frac{1}{2}\ln\left(\frac{5}{3}\right)$
$\frac{1}{3}\ln\left(\frac{5}{3}\right)$
(*Continue, Share*)
(*Continue*, $\alpha>\frac{1}{2}$)
(*Dissolve*, $\alpha<\frac{1}{2}$)
(*Dissolve, Take*)
$\frac{4}{7}\ln\left(\frac{5}{3}\right)$
$\frac{2}{3}\ln\left(\frac{5}{3}\right)$
$\ln\left(\frac{5}{3}\right)$
G

Figure 2 The Pareto-superior equilibrium under complete information.

[27] There is also a continuum of equivalent Pareto-inferior equilibria (*Dissolve, Take*) where the conditional second-order belief of B (which cannot be derived from the correct, unconditonal one) is concentrated on some value above $\alpha^{-}(G,R)$. Since $WS(\alpha;G,R)<0$ for $\alpha>\alpha^{-}(G,R)$, *Take* is a conditional best response.

[28] In the unique *sequential equilibrium* (Battigalli & Dufwenberg 2009), the probability of *Share* is precisely $\alpha^{-}(G,R)$.

Proposition 2 *The Pareto-superior equilibrium of the guilt-reciprocity model (2) under complete information is*

$$\begin{array}{lll} (Continue, Share) & \alpha = \beta = 1 & \textit{if} \quad G + R \geq \ln\left(\frac{5}{3}\right), \\ (Continue, \alpha) & \alpha = \alpha^{-}(G,R) & \textit{if} \quad G + R < \ln\left(\frac{5}{3}\right) \textit{ and } R \geq \frac{1}{2}\ln\left(\frac{5}{3}\right) - \frac{1}{8}G, \\ (Dissolve, \alpha) & \alpha = \alpha^{-}(G,R) & \textit{if} \quad G + R < \ln\left(\frac{5}{3}\right) \textit{ and } \frac{1}{3}\ln\left(\frac{5}{3}\right) < R < \frac{1}{2}\ln\left(\frac{5}{3}\right) - \frac{1}{8}G, \\ (Dissolve, Take) & \alpha = \beta = 0 & \textit{if} \quad G + R < \ln\left(\frac{5}{3}\right) \textit{ and } R \leq \frac{1}{3}\ln\left(\frac{5}{3}\right). \end{array}$$

3.3.2 Incomplete information

A detailed equilibrium analysis of the incomplete-information about psychological preferences requires the specification of players' hierarchies of beliefs about preferences by means of a type structure, which yields a Bayesian psychological game. The analysis of a fully-fledged incomplete-information model is rather complex and beyond the scope of this paper. Here we only provide a qualitative analysis based on intuition.[29]

We start with two observations. First, when subjects are matched at random and do not observe anything, directly or indirectly, about the other subject with which they are matched, the behavior and beliefs of A-subjects must be independent of the behavior, beliefs and psychological utility of B-subjects. Second, in an incomplete-information setting, it is plausible to assume that A-subjects have *heterogeneous* and dispersed beliefs about the psychological utility function u_B. It is even more plausible that B-subjects have heterogeneous and dispersed (second-order) beliefs about such beliefs of the A-subjects. An A-subject is characterized by her hierarchy of exogenous beliefs, where the first-order belief is her belief about u_B; such hierarchy is summarized by A's **type**. A B-subject is characterized by his hierarchy of exogenous beliefs and by his psychological utility u_B; these features are summarized by B's **type,** which comprises B's **utility type**.[30] Hence, from now on, *we describe the equilibrium behavior and beliefs of A-types and B-types.* By the first observation, the types of A and B must be independent; therefore, the beliefs and behavior of A and B must also be independent. We further assume that the utility type of B and his hierarchy of exogenous beliefs are independent.

[29] In the *Online Appendix B* of the paper, at http://www.igier.unibocconi.it/wp506, we present a model consistent with our qualitative analysis. Attanasi *et al.* (2015) provide a Bayesian equilibrium analysis of the Trust Minigame with guilt aversion (but not reciprocity), including the possibility that also the truster A can be guilt averse.

[30] We call "**exogenous**" a belief about an exogenous variable or a parameter. For example, 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. For example, α is an endogenous first-order belief of A, while $\mathbb{E}_B\left(\widetilde{\alpha}\right)$ and $\beta := \mathbb{E}_B[\widetilde{\alpha}|Cont.]$ are both endogenous second-order beliefs of B. Types in the sense of Harsanyi determine only the utility functions and exogenous beliefs.

In our analysis of the complete-information model, we focus on the Pareto-superior equilibrium, which is the one with the highest degree of trust and cooperation. Similarly, here we describe the main features of an incomplete-information equilibrium where a positive fraction of A-types choose *Continue*, which implies that *Continue* has positive probability, so the conditional belief $\mathbb{P}_B(\cdot|Cont.)$ is derived from the initial (unconditional) belief of B.

Since an A-type chooses *Continue* only if she holds a first-order belief $\alpha \geq 1/2$, the conditional second-order belief of a B-type who "virtually observes" *Continue* must assign probability one to $\alpha \geq 1/2$.[31] Note that the same conclusion can be obtained by forward induction (FI), i.e., by assuming that B rationalizes A's choice under the presumption that A is selfish and risk neutral. Essentially, by looking at Bayesian equilibria where a positive fraction of A-types choose *Continue*, we are deriving the common features of equilibria that satisfy this forward-induction requirement (see Attanasi *et al.* 2016). With this, it is appropriate to look at the following **FI-dominance regions** in the space of utility types (G, R) (see Figure 3):

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

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

where $L > \ln(5/3)$ is a commonly known upper bound on parameters.

To see the relevance of these regions, recall that $WS(\alpha; G, R)$ is the willingness to share of B with utility type (G, R) when he is certain that the first-order belief of A is α. Here we take into account that, even if B is not certain about the value of α, he is certain – conditional on *Continue* – that $\alpha \geq 1/2$. Therefore, the B-types with $(G, R) \in \mathbb{S}$ strictly prefer *Share*, and the B-types with $(G, R) \in \mathbb{T}$ strictly prefer *Take*.

This allows us to derive a lower bound and an upper bound on α. Let α_{t_A} and $\mathbb{P}_{t_A}(\mathbb{S})$, respectively, denote the probability assigned by type t_A to *Share* (an endogenous belief determined in equilibrium) and to region $\mathbb{S}$ (an exogenously given belief of t_A). Then $\alpha_{t_A} \geq \mathbb{P}_{t_A}(\mathbb{S})$. Similarly, $1 - \alpha_{t_A} \geq \mathbb{P}_{t_A}(\mathbb{T})$. Of course, since different A-types have different beliefs about B-types, they also hold different first-order beliefs about the strategy of B. But the previous observation allows us to bound first-order beliefs: Let $\underline{\alpha} := \inf_{t_A} \mathbb{P}_{t_A}(\mathbb{S})$, and $\bar{\alpha} := 1 - \inf_{t_A} \mathbb{P}_{t_A}(\mathbb{T})$; then, for every A-type t_A, $\underline{\alpha} \leq \alpha_{t_A} \leq \bar{\alpha}$.

Similarly, different B-types hold different beliefs about A-types, and therefore different

[31] We are considering a "virtual observation" because, in the strategic form of the Trust Minigame, A and B choose strategies simultaneously. Nonetheless, only B's belief conditional on *Continue* is relevant for B's rational choice.

initial beliefs about the choice and first-order beliefs of A, and different conditional beliefs about the first-order beliefs of A. But we can put bounds on these beliefs too: Since $\alpha_{t_A} \geq \mathbb{P}_{t_A}(\mathbb{S})$ (respectively, $1-\alpha_{t_A} \geq \mathbb{P}_{t_A}(\mathbb{T})$), and type t_A chooses *Continue* (respectively, *Dissolve*) if $\alpha_{t_A} > 1/2$ (respectively, $\alpha_{t_A} < 1/2$), it holds that

$$\mathbb{P}_{t_B}\left(\left\{t_A : \mathbb{P}_{t_A}(\mathbb{S}) > \frac{1}{2}\right\}\right) \leq \mathbb{P}_{t_B}\left(Cont.\right) \leq 1 - \mathbb{P}_{t_B}\left(\left\{t_A : \mathbb{P}_{t_A}(\mathbb{T}) > \frac{1}{2}\right\}\right),$$

where $\mathbb{P}_{t_B}$ denotes the belief of a B-type t_B. Furthermore, $\underline{\alpha} \leq \alpha_{t_A} \leq \bar{\alpha}$ imply that initial second-order beliefs satisfy $\underline{\alpha} \leq \mathbb{E}_{t_B}(\widetilde{\alpha}) \leq \bar{\alpha}$.

The behavior of B-types t_B with utility type (G, R) out of the FI-dominance regions depends on their equilibrium conditional belief $\mathbb{P}_{t_B}\left(\cdot|Cont.\right)$.[32] Since function $WS(\alpha; G, R)$ is increasing on $[1/2, 1]$ iff $R/G \leq 1/2$ (see (8)), for every utility type (G, R) with $G \geq 2R$, a "higher" conditional second-order belief (in the sense of stochastic dominance) implies a higher willingness to share $\mathbb{E}_{t_B}[WS(\widetilde{\alpha}; G, R)|Cont.]$.

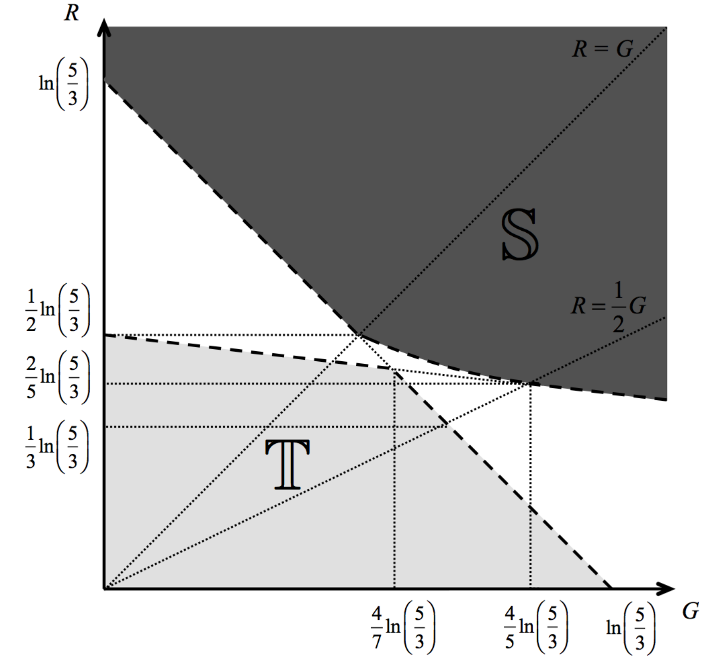

R
$\ln\left(\frac{5}{3}\right)$
$R = G$
$\mathbb{S}$
$\frac{1}{2}\ln\left(\frac{5}{3}\right)$
$R = \frac{1}{2}G$
$\frac{2}{5}\ln\left(\frac{5}{3}\right)$
$\frac{1}{3}\ln\left(\frac{5}{3}\right)$
$\mathbb{T}$
$\frac{4}{7}\ln\left(\frac{5}{3}\right)$
$\frac{4}{5}\ln\left(\frac{5}{3}\right)$
$\ln\left(\frac{5}{3}\right)$
G

Figure 3 FI-dominance regions for *Share* and *Take* when B is certain that $\alpha \geq 1/2$.

Given reasonable assumptions about the statistical distribution of A-types and B-types, our analysis yields the following qualitative results:[33]

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

[33] See the results derived in *Online Appendix B* given the assumptions about the distribution of types.

Proposition 3 *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*; the initial first-order beliefs about α are bounded by $\underline{\alpha}$ and $\bar{\alpha}$: For every t_A, $\underline{\alpha} \leq \alpha_{t_A} \leq \bar{\alpha}$.*
(2) [B-heterogeneity] B*-types have heterogeneous, dispersed initial beliefs about* A*'s strategy and α; the unconditional second-order beliefs about α are bounded by $\underline{\alpha}$ and $\bar{\alpha}$: For every t_B, $\underline{\alpha} \leq \mathbb{E}_{t_B}(\widetilde{\alpha}) \leq \bar{\alpha}$. 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, utility type, and beliefs of* B*;*
(3.ii) [Independence within roles] B*'s first- and second-order beliefs are independent of the utility 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 $G > 2R$ who choose* Share *is positively correlated with the conditional second-order belief.*

3.3.3 Theoretical predictions and experimental design

The theoretical analysis in 3.3.1 (complete information) and 3.3.2 (incomplete information) leads to several testable predictions. All these predictions refer to B's utility 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 utility 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 utility type, 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 fall into two categories.

First, we check whether B-subjects have belief-dependent preferences and how they interact with the information structure in each phase-treatment combination:

- In the *complete-information phase* (phase 3 of *QD*), we predict a polarization of behavior and beliefs because common knowledge of B's utility type works as a coordination device. In fact, if B is sufficiently selfish (low guilt and reciprocity parameters), *Dissolve* is the unique equilibrium outcome and the probability of *Share* (which is equal to the first-order belief α) is lower than 1/2. In the complementary region of the parameter space, *Continue* is the Pareto-superior equilibrium strategy of A, and the probability of *Share* is higher than 1/2 (see Proposition 2 and Figure 2).

- In the *incomplete-information phases* (all other phase-treatment combinations), there are more heterogeneity of behavior and more dispersed beliefs, including "intermediate" beliefs. This is quite obvious for A-subjects (assuming heterogeneous, dispersed beliefs about u_B). More interestingly, there is a parameter region with intermediate values of G (and low values of R) where B-subjects would cooperate and hold high second-order beliefs under complete information, while they exhibit less cooperative behavior and intermediate second-order beliefs under incomplete information (see Proposition 3, compare Figure 3 with Figure 2).

Second, we qualitatively compare players' behavior and beliefs across treatments and across phases of the same treatment:

- *QD* vs. *NoQ* and *QnoD*: *Between* the "complete-information treatment" *QD* and the "incomplete-information treatments" *NoQ* and *QnoD*, we predict differences in the direction of a polarization of behavior and beliefs in phase 3 of *QD*, and no differences in phase 1.

- Phase 3 *vs.* Phase 1: *Within QD*, we predict differences in the direction of a polarization of behavior and beliefs in phase 3. On the other hand, we expect no differences between phases 1 and 3 in *NoQ* and *QnoD*.

In Section 4, we discuss the data guided by the theoretical predictions for the two different information regimes.[34]

[34] As specified above, due to the multiplicity of equilibria, the comparative statics of the two types of information in part rely on our selection criterion.

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, in 4.1 we present the classification of B's belief-dependent preferences derived from the answers to the questionnaire of Table 3. With this classification in mind, we analyze A's and B's behavior in the Trust Minigame using the equilibrium predictions of Section 3.3. In particular, in 4.2 we use the complete-information predictions to analyze subjects' behavior in phase 3 of the treatment with questionnaire disclosure (QD). In 4.3 we use our qualitative, incomplete-information predictions to analyze behavior in phase 1 of QD and in the treatments without questionnaire disclosure (NoQ and $QnoD$). In 4.4, we compare behavior in all these phase-treatment combinations with behavior in phase 3 of QD.

4.1 Experimental Elicitation of Belief-Dependent Preferences

As explained in Section 3.2, the experimental elicitation of B's belief-dependent preferences in the Trust Minigame relies on the questionnaire of Table 3. Here, we analyze the answers of each B-subject to the questionnaire. We call "**payback pattern**" the actual answers of a B-subject, one payback value for each hypothesized α (A's belief about B's strategy). Our aim is to estimate, for each B-subject, the triple (G, R, p) that identifies B's best response to the 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.

Recall that the payback pattern gives only 11 observations for B's payback function, i.e. one for each $\alpha \in \{0, 10\%, ..., 100\%\}$. 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 (least squared error) 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.[35]

We find significant heterogeneity around the estimated averages: the estimated variances of $\hat{G}$, $\hat{R}$ and $\hat{p}$ are all significantly different from zero.[36] In Table 4, we report the distribution

[35] In the *Online Appendix C* of the paper, at http://www.igier.unibocconi.it/wp506, we provide raw data and the non-linear least square estimates $\hat{G}$, $\hat{R}$ and $\hat{p}$ (with associated standard deviations) for the 160 B-subjects in our experiment.

[36] In particular, 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.

of B-subjects' estimated utility types across the possible shapes of the corresponding payback function $\xi(\alpha)$. According to Proposition 1, the theoretical payback function $\xi(\alpha)$ can be (i) *increasing* ($G > R$ and R low), (ii) *constant* ($G = R$ and R low), (iii) *first decreasing and then increasing* ($G > R$ and intermediate R), and (iv) *decreasing* ($R > G$, or R high). We consider separately the case $G = R = 0$ of *selfish* preferences. This explains the categorization of Table 4.[37]

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

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

(*) Column *NoQ-QnoD* pools the observations of *NoQ* and *QnoD*.

Considering the actual answers to the questionnaire (the payback pattern), rather than the estimated payback function, we find 138/160 (86%) B-subjects whose payback pattern mimics one of the quasi-convex shapes of $\xi(\alpha)$ implied by our model. We use the estimated parameters $\hat{G}$ and $\hat{R}$ to classify also the remaining 22/160 (14%) B-subjects within one of the five categories of Table 4 (for a similar method, see Costa-Gomes *et al.* 2001).

Since we find no significant within-treatment difference in the distribution of estimated utility types,[38] we pool the data within each treatment. (Note that for treatment *NoQ* we rely on the final questionnaire – the only one filled-in by B-subjects in this treatment –,

[37] Notice that $\hat{p}$ 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 subject, but not in the analysis of the Trust Minigame.

[38] To check for framing effects, in half of the sessions of each treatment, the 11 lines of the questionnaire of Table 3 have been shown in *reverse order*, starting with 100% instead of 0%. We find no significant order effect.

while for treatments *QnoD* and *QD* we refer to the questionnaire in phase 2 – the first one filled-in in these treatments.) Furthermore, we find 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*). We also find no significant difference between the distributions of utility types in *NoQ-QNoD* and *QD* (last two columns of Table 4: χ^2 test, *P-value* = 0.734). This means that the presence or absence of information disclosure essentially does not affect subjects' answers to the questionnaire. This interpretation is further supported by the fact that – with very few exceptions – B-subjects in the *QD* treatment did not change their payback pattern from the phase 2 questionnaire to the final questionnaire (cf. Table 2).[39]

Table 4 shows that, independently of the treatment, the guilt component of psychological utility function u_B is prevalent for more than half of the B-subjects,[40] while reciprocity prevails for only 16% of them. There is, however, 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 ("balanced") estimated payback function (payback independent of A's first-order belief). 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.

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 sub-populations of subjects within the different treatments.

[39] 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 classification in the categories of Table 4. We acknowledge that this tendency of B-subjects to provide the same answers to the (similar) questions in phase 2 and the final questionnaire could be due to a consistency motive (see, e.g., Podsakoff *et al.* 2003).

[40] In a trust game similar to the one in Charness and Dufwenberg (2006), Ederer & Stremitzer (2015) find that more than half of the trustees exhibit guilt aversion. In a series of dictator games with various stake sizes, Bellemare *et al.* (2015) find that 41.6% of dictators are guilt averse.

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 parameters $\hat{G}$ and $\hat{R}$ obtained from the payback pattern (predicted behavior). Second, we compare actual behavior with predicted behavior, at the pair and individual level.[41]

Figure 4 reports the *actual 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. The figure refers to the four regions of the parameter space (G, R) of Figure 2, which correspond to the complete-information predictions of Proposition 2.[42]

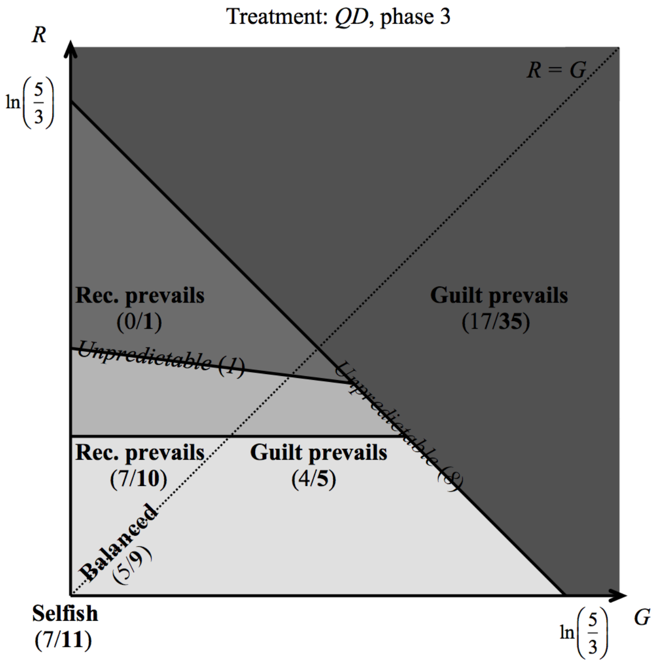


Figure 4 Actual *vs.* complete-information predicted behavior (strategy pairs) in phase 3 of QD.

In each region and for each category of utility type from Table 4 (guilt prevails, reciprocity prevails, etc.), we report in bold the number of "**predictable**" B-subjects, i.e. those for whom we obtain a clear complete-information prediction for the corresponding matched pairs;[43] we report in Italics the number of remaining ("*unpredictable*") subjects. Before the

[41] We implemented a stranger-matching design: A-subjects and B-subjects are randomly re-matched so as to have different pairs in phase 1 and in phase 3 and avoid repeated-game effects. With the goal of providing a clean check of within-treatment differences, throughout Section 4 we analyze pairs' behavior in phase 1 of each treatment according to the random matching of phase 3. This can be done at no cost, since A's (B's) choice in phase 1 is told to the B (A) matched with her (him) only at the end of the experiment.

[42] Recall from Figure 2: the lightest color corresponds to (*Dissolve*,*Take*), the darkest color corresponds to (*Continue*,*Share*), etc.

[43] For these B-subjects, the estimated utility type $(\hat{G}, \hat{R})$ can be assigned to one of the four regions of the

number of predictable B-subjects in *QD*, we report the number of the corresponding matched pairs who behave as predicted in phase 3 of *QD* (regular character).

Pairs' predicted behavior Given the estimated utility type $(\hat{G}, \hat{R})$, we can make a prediction (*Share vs. Take*) for about 89% (71/80) of B-subjects in *QD* (bold numbers in Figure 4). All but one of the corresponding matched pairs belong to either the (*Dissolve*,*Take*) region or the (*Continue*,*Share*) region. The (*Dissolve*,*Take*) region includes all pairs with a "balanced" B-subject, all but one predictable pairs with a B-subject for whom reciprocity prevails 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, 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 2. 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 4 as **"high-guilt" types** (35/71) and to the B-subjects in the (*Dissolve*,*Take*) region of Figure 4 as **"low-guilt" types** (35/71). The latter subgroup includes all but one predictable B-subjects with $\hat{R} > \hat{G}$ and all those with $\hat{G} = \hat{R} \geq 0$.[44]

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* with probability 1 under complete information are "high-guilt" types, and no B-subject for whom reciprocity prevails is predicted to choose *Share* with probability 1 under complete information. All this holds independently of the treatment.

Actual behavior of matched pairs We use as controls the phase-treatment combinations where the filled-in questionnaire is not disclosed, henceforth **"incomplete-information phases"** (phase 1 of *QD*, phases 1 and 3 of *NoQ-QnoD*). Throughout the paper we provide

parameter space (G, R) of Figure 2 (complete-information predictions) with a level of significance of at most 10% (P-values estimated by bootstrap).

[44] We replicated the above exercise also for the 80 B-subjects in *NoQ-QnoD*: The parameter constellations estimated in *NoQ-QnoD* lead to similar complete-information predicted behavior as in *QD* under the counterfactual assumption that subjects have complete information also in phase 3 of *NoQ-QnoD* (compare bold numbers in Figure 4 and in Figure 4ter in the *Online Appendix C* of the paper, at http://www.igier.unibocconi.it/wp506). The absence of such treatment bias enhances the portability of our methodology (see Camerer & Ho 1999).

aggregate results about these phase-treatment combinations, because we do not find significant between-treatment, or within-treatment differences.

In Figure 4 the ratios of actual (regular character) *vs.* predicted (bold) behavior in phase 3 of *QD* show a 56% (40/71) rate of success of the complete-information predictions for phase 3 of *QD*. This is significantly higher (χ^2 test, *P-value* = 0.011) than for the incomplete-information phases (39% on average over the incomplete-information phases, 74/191).[45]

Our complete-information predictions are particularly successful for pairs being predicted to choose (*Dissolve*,*Take*): 66% (23/35) in phase 3 of *QD*. However, the greatest difference with respect to the controls is found for pairs being predicted to choose (*Continue*,*Share*): 47% (17/36) in phase 3 of *QD vs.* 14% (12/86) in the incomplete-information phases, significant at the 1% level.

The following result summarizes the main experimental findings about matched pairs' actual *vs.* predicted behavior under complete information.

Result 3 Complete-information predictions are able to explain 56% of actual strategy pairs after questionnaire disclosure (phase 3 of treatment *QD*). In particular, 66% of pairs in the (*Dissolve*,*Take*) region and 47% of pairs in the (*Continue*,*Share*) region behave as predicted.

In Figure 5 we deepen the analysis presented in Figure 4. We present *subjects' actual choices and beliefs* in phase 3 of *QD*, disentangled by role and by B's utility type. Given that there is no predictable utility type in region ($Dissolve, \alpha^- < 1/2$) and just one predictable utility type in region ($Continue, \alpha^- > 1/2$) of Figure 4, here we focus only on the two regions of polarized predictions (*Continue*,*Share*) and (*Dissolve*,*Take*), with 35 subjects in each region, respectively "high-guilt" types and "low-guilt" types. First, we discuss experimental results about *choices and first-order beliefs* of A-subjects in phase 3 of *QD*.[46] Then, we discuss experimental results about *choices*, *first- and second-order beliefs* of B-subjects in phase 3 of *QD*.[47]

[45] The corresponding figures for the incomplete-information phases can be found in the *Online Appendix C* of the paper, at http://www.igier.unibocconi.it/wp506.

[46] 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.

[47] 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

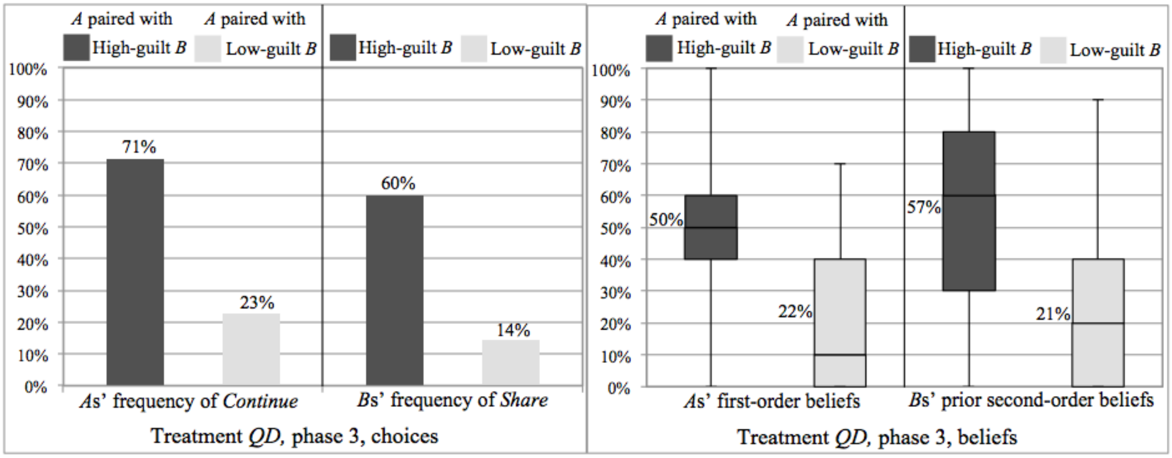


Figure 5 *A*s and *B*s' choices and beliefs in phase 3 of *QD*, disentangled by *B*'s type.

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

As' actual behavior and beliefs As reported in Figure 5, *A*-subjects matched with a high-guilt *B*-subject show a significantly greater frequency of *Continue* (+49%, χ^2 test: *P-value* $= 0.000$) and a significantly greater first-order belief α (+28% on average, Mann-Whitney test: *P-value* $= 0.000$). A significant (at the 1% level) positive correlation is found between the *Continue* choice and α (Spearman's $\rho = 0.57$, *P-value* $= 0.000$).

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 utility type $(\hat{G}, \hat{R})$ relevant for the equilibrium analysis of Proposition 2 – and both *A*'s choice of *Continue* ($\rho = 0.45$) and α ($\rho = 0.43$). This is mainly due to the guilt component $\hat{G}$ ($\rho = 0.46$ with *Continue*; $\rho = 0.47$ with α), while for $\hat{R}$ we find a low negative correlation with both *A*s' choice and belief.[48]

Finally, if we disentangle the *A*-subjects in phase 3 of *QD* according to the matched (estimated) utility type – high-guilt *vs.* low-guilt – and we focus on each subgroup separately,

is less polarized than the true one (Charness & Dufwenberg 2006 face 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.

[48] We verified that $\hat{G}$ and $\hat{R}$ are statistically independent ($\rho = -0.10$, *P-value* $= 0.191$). This allows us to run the correlation analysis with *A*s' choice and first-order belief for $\hat{G}$ and $\hat{R}$ separately.

we find no significant correlation between $(\hat{G}, \hat{R})$ and both the *Continue* choice and α. This is in line with the complete-information predictions: all A-subjects in the former (latter) subgroup choose *Continue* (*Dissolve*) if B's utility type $G + R > \ln(5/3)$, the threshold for (*Continue*,*Share*) in Proposition 2.

The following result summarizes the more salient experimental findings about A-subjects' actual *vs.* predicted 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 greater for A-subjects matched with high-guilt B-subjects. More generally, both trust and A's first-order beliefs are positively correlated with the disclosed guilt type of B.

Bs' actual behavior and beliefs As reported in Figure 5, high-guilt B-subjects show a greater frequency of *Share* (+46%) and greater unconditional second-order beliefs $\mathbb{E}_B(\widetilde{\alpha})$ (+36% on average) than low-guilt B-subjects, both differences being significant at the 1% level.

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, and with an abuse of language, we refer to such reported beliefs as Bs' **first-order point-belief**. With this, we find a strongly significant positive correlation between *Share* and the first-order point-belief ($\rho = 0.45$, *P-value* $= 0.000$).

The correlation between *Share* and $\mathbb{E}_B(\widetilde{\alpha})$ is again strongly significant ($\rho = 0.58$, *P-value* $= 0.000$), and significantly higher ($\rho = 0.70$) 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 find that 95% (18/19) 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 ($\rho = 0.44$), first-order point belief ($\rho = 0.53$), and $\mathbb{E}_B(\widetilde{\alpha})$ ($\rho = 0.49$) are consistent with the theoretical predictions. In particular, since B is aware that his utility type $(\hat{G}, \hat{R})$ is disclosed to A, his beliefs about A's behavior and beliefs move with $\hat{G} + \hat{R}$.[49]

Disentangling by utility type – high-guilt *vs.* low-guilt –, we find no significant correlation between B's disclosed utility type $(\hat{G}, \hat{R})$ and B-subjects' choices and first- and second-order beliefs, in any of the two sub-groups considered separately. This confirms the complete-information predictions: B's choice depends on whether G+R is above or below the threshold

[49] 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 (non-significant) negative correlation with $\hat{R}$.

in Proposition 2, but not on their precise value.
The following result summarizes the more salient experimental findings about B-subjects' actual *vs.* predicted 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 greater 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 utility type. To check the 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 3:

(1) A-heterogeneity A-subjects' first-order beliefs are heterogeneous and dispersed: Only 22% (1%) of A-subjects have $\alpha = 0$ ($\alpha = 1$), the coefficient of variation of α being 0.90. We also find a significant difference (at the 1% level) in the frequency of *Continue* choices (85% *vs.* 12%) 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 40% (60%) of B-subjects reporting *Continue* (*Dissolve*). The unconditional second-order beliefs are heterogeneous and dispersed: Only 26% (3%) of B-subjects have $\mathbb{E}_B(\widetilde{\alpha}) = 0$ ($\mathbb{E}_B(\widetilde{\alpha}) = 1$), the coefficient of variation of $\mathbb{E}_B(\widetilde{\alpha})$ being 0.89. 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 almost all of them (95%) have $\mathbb{E}_B(\widetilde{\alpha}) > 0$, but only 46% of them have $\mathbb{E}_B(\widetilde{\alpha}) \geq 1/2$ (average belief 0.68).

(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, estimated utility type, and first- and second-order unconditional beliefs ($\max|\rho| = 0.07$ in the four rank correlations). A similar result holds for A's first-order belief ($\max|\rho| = 0.05$).

(3.ii) Independence within roles B's second-order beliefs are independent ($\rho = 0.10$, *P-value* $= 0.119$) of the estimated utility type $(\hat{G}, \hat{R})$, corroborating our assumption that the epistemic component of B's type is independent of the utility component.[50] We find a small correlation ($\rho = 0.15$, *P-value* $= 0.019$) between B's first-order beliefs and $(\hat{G}, \hat{R})$.

(4) FI-dominance We organize B-subjects' choices according to the incomplete-information predictions using $\hat{G}$ and $\hat{R}$ obtained from the payback pattern (predicted behavior). Figure 6 refers to the three regions of predictions in the parameter space (G, R) of Figure 3. This figure has been constructed using the same method and notation as Figure 4. On the left panel of Figure 6, we report actual (regular character) *vs.* predicted (bold) behavior in phase 1 of *QD*; on the right panel, we report actual *vs.* predicted behavior in phases 1 and 3 of *NoQ-QnoD*.

We find 65/80 "predictable" B-subjects in *QD* and 50/80 "predictable" B-subjects in *NoQ-QnoD*.[51] In the following analysis, we will mainly focus on the two FI-dominance regions, i.e. $\mathbb{T}$ (light-grey color) and $\mathbb{S}$ (dark-grey color). Similarly to the complete-information case, given the categories of elicited utility types in the two regions, we refer to B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{T}$ as "low-guilt" types and to B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ as "high-guilt" types. Indeed, the former coincide with low-guilt types under complete information, while the latter are a subset of high-guilt types under complete information.[52]

Relying on the incomplete-information predictions in Figure 6, and considering together the incomplete-information phases in *QD* and in *NoQ-QnoD*, we find that *Share* is chosen by 42% of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$, while it is chosen by only 15% of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{T}$, the difference being significant at the 1% level; also $\mathbb{E}_B(\widetilde{\alpha})$ is significantly (at the 5% level) greater for high-guilt than for low-guilt types. The result that less than half of B-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ choose *Share* can be explained by the fact that for many of them the forward-induction assumption that $\beta \geq 1/2$ does not seem to hold (see Attanasi *et al.* 2016 for a theoretical explanation of this result).[53] 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 89%

[50] See the theoretical model in the *Online Appendix B*, at http://www.igier.unibocconi.it/wp506. We make the assumption also there.

[51] "Predictable" B-subjects in Figure 6 are those for whom the estimated utility type $(\hat{G}, \hat{R})$ can be assigned to one of the three regions of the parameter space (G, R) of Figure 3 (incomplete-information predictions) with a level of significance of at most 10% (P-values estimated by bootstrap).

[52] Compare the left panel of Figure 6 with Figure 4, and the right panel of Figure 6 with Figure 4ter or Figure 4quater in *Online Appendix C*.

[53] Attanasi *et al.* (2016) analyze a model where A, like B, may be guilt averse. When A is perceived by B as potentially guilt averse, action *Continue* may be interpreted as a desire not to disappoint B, hence it may well be the case that $\beta < 1/2$. 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.

of those with $(\hat{G}, \hat{R}) \in \mathbb{S}$ and $\mathbb{E}_B(\widetilde{\alpha}) \geq 1/2$ choose *Share*.

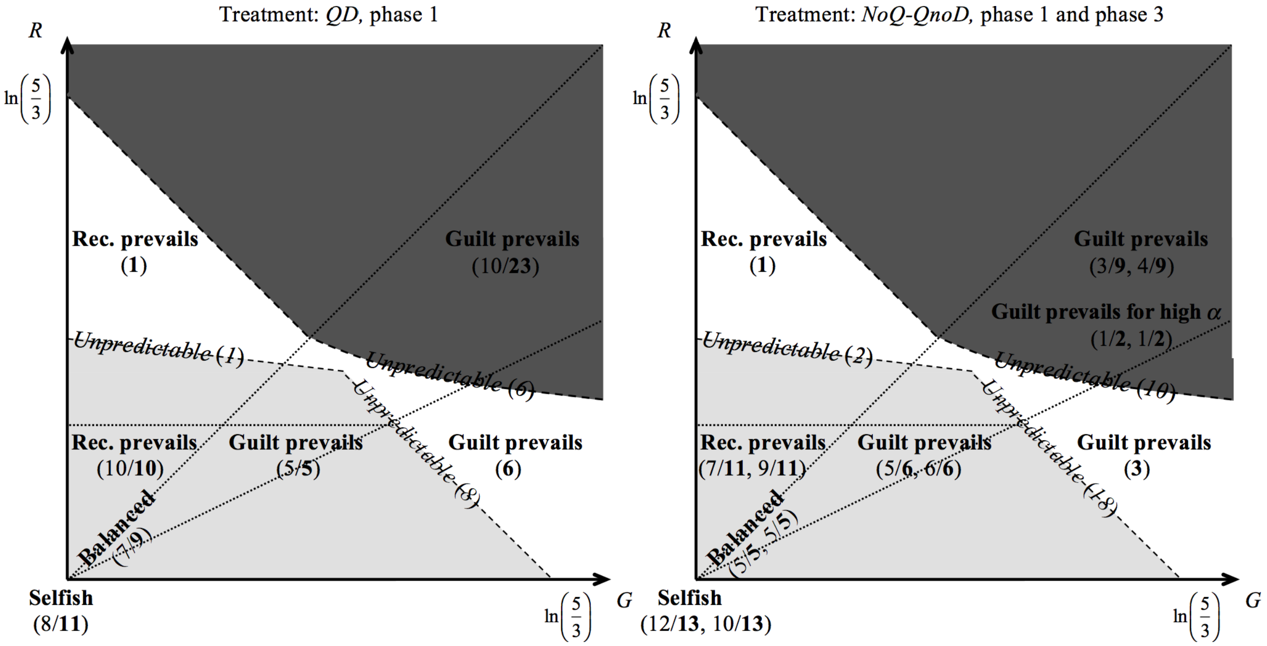


Figure 6 Actual *vs.* incomplete-information predicted behavior of B-subjects.

The figure reports: for *QD* and *NoQ-QnoD* separately, in **bold**, the number of B-subjects for whom we obtain a clear incomplete-information prediction ("predictable" B-subjects) in each region (from Figure 3), and the category to which they belong (from Table 4); in *Italics*, the number of remaining subjects ("unpredictable" B-subjects). The left panel refers to phase 1 of *QD*; each ratio indicates actual (regular character) *vs.* predicted (bold) behavior. The right panel refers to *NoQ-QnoD*; the first ratio refers to actual *vs.* predicted behavior in phase 1, the second ratio refers to actual *vs.* predicted behavior in phase 3.

(5) Choice-belief correlation The proportion of intermediate types (region $(\mathbb{S} \cup \mathbb{T})^c$) choosing *Share* is between the corresponding proportions for types in region $\mathbb{T}$ and in region $\mathbb{S}$ (31%, significantly different from the other two proportions at the 1% level). Focusing only on those intermediate types with *Continue* as first-order point-belief, this proportion slightly increases (36%), and we find a positive correlation ($\rho = 0.49$) between the *Share* choice and $\mathbb{E}_B(\widetilde{\alpha})$, a rough estimate of β. A significant positive correlation ($\rho = 0.33$, *P-value* $= 0.043$) is found considering *all* the predictable B-subjects with *Continue* as first-order point-belief and $\hat{G} > 2\hat{R}$.

The following result summarizes the more salient experimental findings about matched pairs' actual *vs.* predicted behavior and beliefs under incomplete information.

Result 6 In line with the incomplete-information qualitative 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. Furthermore, B-subjects predicted to choose *Share* actually do it much more frequently than those predicted to choose *Take* (42% *vs.* 15%). For B-subjects with $\hat{G} > 2\hat{R}$ who expect *Continue*, the *Share* choice is positively correlated with the belief about α.

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. First, we compare frequencies of strategy profiles chosen by complete-information predictable pairs in phase 3 of *QD vs.* the phase-treatment combinations without disclosure (incomplete-information phases). Next, we rely on the separation between high-guilt and low-guilt types introduced in the complete-information case (see Table 4 and Figure 5) to make within- and between-treatment comparisons of Bs' and matched As' behavior and beliefs.

Pairs' actual behavior Table 5 reports the frequencies of actual strategy profiles across *all* the complete-information predictable pairs, in phase 3 of *QD* and in the pooled incomplete-information phases. In the former, in line with the complete-information predictions, there is a significant correlation between *Continue* (resp. *Dissolve*) and *Share* (resp. *Take*) in phase 3 of *QD* ($\rho = 0.35$, *P-value* $= 0.003$). A significant correlation (Pearson's $r = 0.34$, *P-value* $= 0.004$) is also found between the elicited values of α and $\mathbb{E}_B(\widetilde{\alpha})$ for the complete-information predictable pairs in *QD*. In the incomplete-information phases, as expected in a random-matching setting, both the choices and the beliefs of A and B are independent (indeed, we find $\rho = -0.02$ and $r = 0.002$, respectively).

	Take	*Share*	
Diss.	42% (30/71)	11% (8/71)	53%
Cont.	21% (15/71)	25% (18/71)	46%
	63%	36%	

Phase 3 of *QD*

	Take	*Share*	
Diss.	53% (102/191)	20% (38/191)	73%
Cont.	20% (38/191)	7% (13/191)	27%
	73%	27%	

Pooled incomplete-information phases

Table 5 Frequencies of actual 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": a high concentration of pairs in the region of trust & cooperation and in the region of no-trust & selfish behavior. Our interpretation of this result is the following. 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 utility type to A. When A receives the filled-in questionnaire of a high-guilt (resp., low-guilt) type, she tends to believe that B would choose *Share* (resp., *Take*) and therefore she also 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 2 in mind, we extend Figure 5 (which only refers to phase 3 of *QD*) and present actual choices and beliefs of pairs with a high-guilt *vs.* a low-guilt type, disentangled by role and phase-treatment combination. Figure 7 reports As' actual frequency of *Continue* choices and first-order belief (α), while Figure 8 reports Bs' actual frequency of *Share* choices and unconditional second-order belief ($\mathbb{E}_B(\widetilde{\alpha})$).

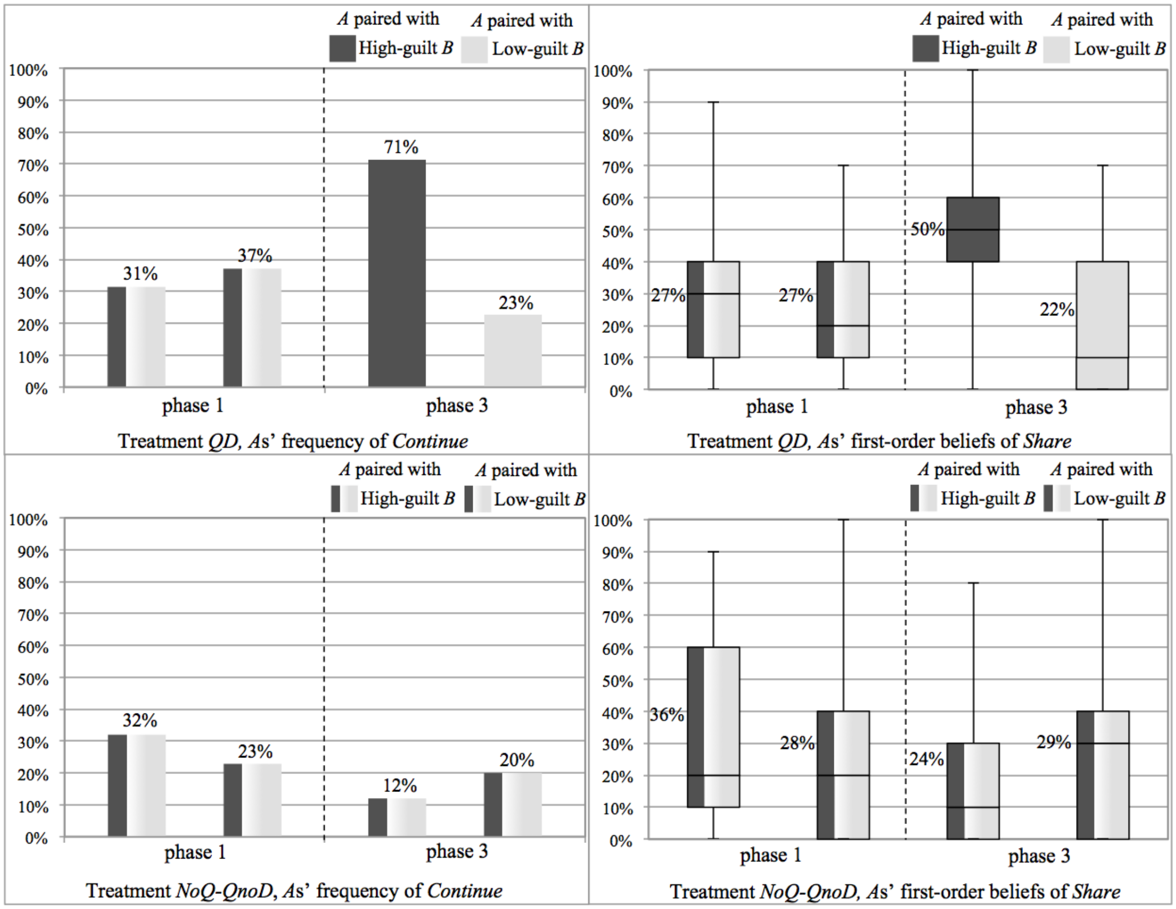


Figure 7 As' freq. of *Continue* and first-order beliefs, by treatment, phase and matched B's type.

For phase 3 of *QD*, histograms of choices (left panels) and box-plots of beliefs (right panels) coincide with those of Figure 5, and the same color code is used. For As in the incomplete-information phases, we use a mixture of the three colors of Figure 3 to emphasize that these subjects do not know the utility type of the B-subject they are matched with.

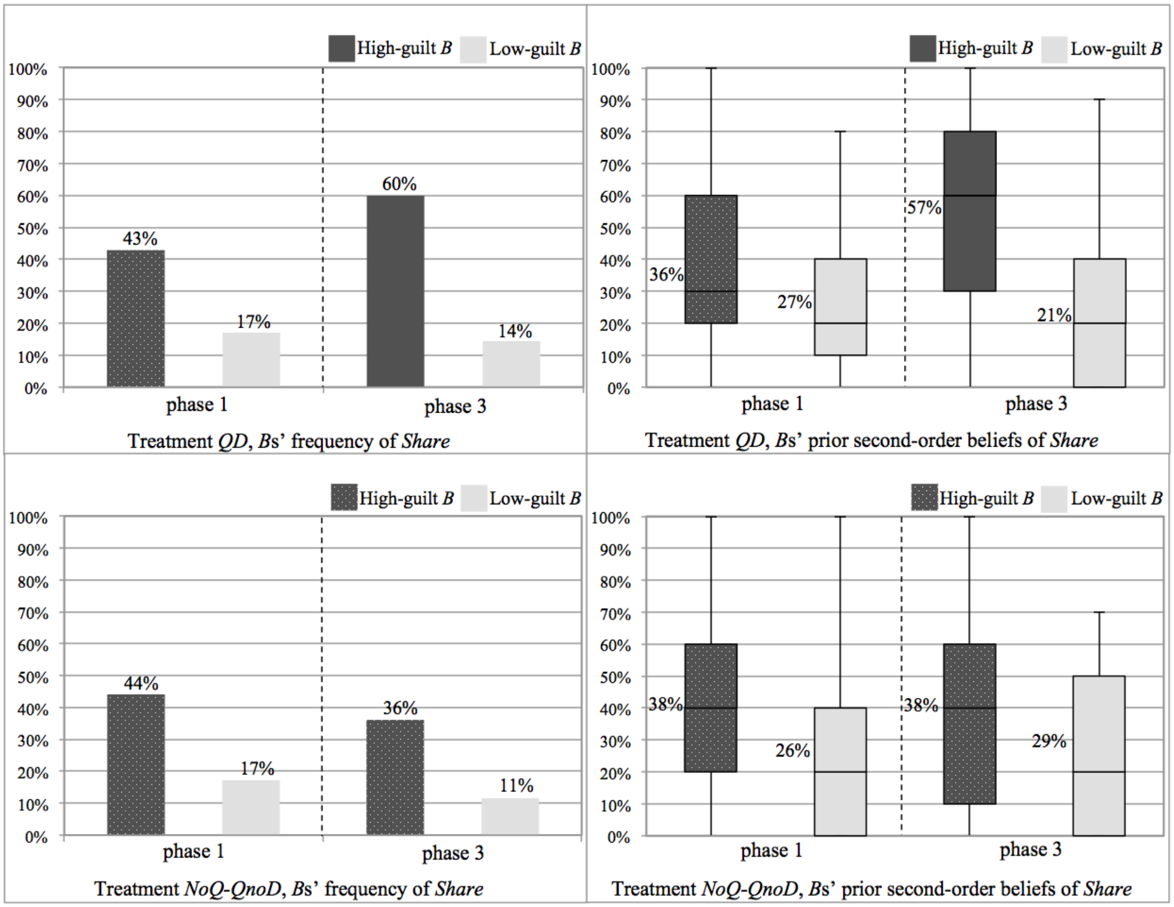


Figure 8 Bs' freq. of *Share* and initial second-order beliefs, by treatment, phase and B's type.

For phase 3 of QD, histograms of choices (left panels) and box-plots of beliefs (right panels) coincide with those of Figure 5, and the same color code is used. For high-guilt Bs, the dark-grey color is meshed with small white dots to indicate that many of them have $(\hat{G},\hat{R})$ in the *Share* region of Figure 3, while a few of them belong to the intermediate white-colored region in Figure 3, or are unpredictable under incomplete information (compare Figure 4 with Figure 6). All low-guilt Bs belong to the *Take* region of Figure 3, hence we use the corresponding light-grey color.

As' actual behavior and beliefs 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.

Within-treatment and between-treatment comparisons work very well for A-subjects matched with a *high-guilt* type: Within treatment, we find a significantly (at the 1% level) greater frequency of *Continue* (+40%) and α (+23% on average) in phase 3 than in phase 1 of QD. Between treatments, we find a similar result by comparing phase 3 of QD to phase 3 of

NoQ-QnoD: respectively, +58% and +26% on average, both significant at 1%.
Within-treatment and between-treatment comparisons are less striking for A-subjects matched with a *low-guilt* type: The decrease from phase 1 to phase 3 of the frequency of *Continue* (−14%) and of α (−5%) within *QD* is not significant, and no significant difference is found (+3% for *Continue* and −7% for α) by comparing phase 3 between *QD* and *NoQ-QnoD*.

Bs' actual behavior and beliefs Within-treatment and between-treatment comparisons work quite well for *high-guilt* B-subjects: We find a higher frequency of *Share* (+17%, *P-value* = 0.151) and significantly greater second-order beliefs (+34% on average, *P-value* = 0.005) in phase 3 than in phase 1 of *QD*.[54] Between treatments, we find similar differences, both significant, by comparing phase 3 of *QD* to phase 3 of *NoQ-QnoD*: +24% (*P-value* = 0.067) for the frequency of *Share*, and +19% on average (*P-value* = 0.028) for $\mathbb{E}_B(\widetilde{\alpha})$.
Within-treatment and between-treatment comparisons work well also for *low-guilt* B-subjects: The predicted behavior is the same under both complete and incomplete information (B-subjects in the (*Dissolve*,*Take*) region in Figure 2 also have $(\hat{G}, \hat{R}) \in \mathbb{T}$ in Figure 3), 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 to phase 1 within *QD*, or phase 3 between *QD* and *NoQ-QnoD*.

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

Result 7 Polarization of subjects' behavior and beliefs due to 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 Brief summary of findings

To sum up, our theoretical analysis of B's answers to the questionnaire (Proposition 1) is able to capture the great majority of B-subjects' payback patterns, most of them being belief-dependent (Result 1). Among these belief-dependent types, we find that all B-subjects predicted to share under complete information (Proposition 2) are "high-guilt" types, i.e. types for whom the other-regarding attitude $G + R$ is above the theoretical threshold, and

[54] The difference in the frequency of *Share* should be caused by the high-guilt subjects that do not belong to the FI-dominance region $\mathbb{S}$. However – compare Figure 4 with the left-hand side of Figure 6 –, there are few such subjects (12); therefore, we are not surprised that this difference is not significant.

the guilt component G is large in absolute terms and relative to the reciprocity component R (Result 2).

Our theoretical model does not predict very accurately the behavior of estimated utility types, but it captures well the central tendencies of the data (Result 3 for the complete-information phase, and Result 6 for incomplete-information phases): The theory summarized in Proposition 2 implies the polarization of behavior and beliefs found only in phase 3 of the *QD* (questionnaire disclosure) treatment, where the utility 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 (Proposition 3).

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 significantly greater under questionnaire disclosure (Result 7), as predicted by the theory (Propositions 2 and 3).

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.

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*, ed. by A. Innocenti and P. Sbriglia. 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*, in print.

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

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.

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 (2015): "Heterogeneous Guilt Aversion and Incentive Effects," Mimeo, Tilburg University.

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.

Camerer, C., and T. H. Ho (1999): "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67, 827–874.

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.

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.

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

Cox, J. C. (2004): "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46, 260–281.

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).

Dufwenberg, M. (2002): "Marital Investment, Time Consistency and Emotions," *Journal of Economic Behavior and 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.

Ederer, F., and A. Stremitzer (2015): "Promises and Expectations," Cowles Foundation Discussion Paper No. 1931.

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990–1008.

Ellingsen, T., M. Johannesson, S. Tjotta, and G. Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior*, 68, 95–107.

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

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

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

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.

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.

Kawagoe, T., and Y. Narita (2014): "Guilt Aversion Revisited: An Experimental Test of a New Model," *Journal of Economic Behavior and Organization*, 102, 1–9.

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

Kuhnen, C., and A. Tymula (2012): "Feedback, Self-esteem and Performance in Organizations," *Management Science*, 58, 94–113.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., 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.

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.

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

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

Siniscalchi, M. (2014): "Sequential Preferences and Sequential Rationality," Mimeo, Northwestern University.

Stanca, L., L. Bruni, and L. Corazzini (2009): "Testing Theories of Reciprocity: Do Motivations Matter?" *Journal of Economic Behavior & Organization*, 71, 233–245.

Tadelis, S. (2011): "The Power of Shame and the Rationality of Trust," Mimeo, UC Berkeley.

Toussaert, S. (2015): "Intention-Based Reciprocity and Signalling of Intentions," Mimeo, New York University.

Vanberg, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica,* 76, 1467–1480.