Economic Theory
https://doi.org/10.1007/s00199-025-01645-5

RESEARCH ARTICLE

Disclosure of belief–dependent preferences in a trust game

Giuseppe Attanasi[1,5] · **Pierpaolo Battigalli**[2] · **Elena Manzoni**[3] · **Rosemarie Nagel**[4]

Received: 31 August 2022 / Accepted: 1 February 2025

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

Keywords Experiments · Trust game · Guilt · Reciprocity · Complete information · Incomplete information

We thank Olivier Armantier for great support in the statistical analysis. We thank for useful discussions and comments Chiara Aina, Carlo Andreatta, Stefania Bortolotti, Roberto Corrao, Martin Dufwenberg, Samuele Dotta, Alejandro Francetich, Pierfrancesco Guarino, Andrea Guido, Sem Manna, Sara Negrelli, Salvatore Nunnari, Fabrizio Panebianco, Jacopo Perego, Ariel Rubinstein, Viola Sigismondi, Nicolas Sourisseau, Fabio Tufano, an anonymous referee, an anonymous Associate Editor and the Guest Editor David Cooper. G. Attanasi gratefully acknowledges financial support by the ERC (Grant DU283953), "Attractivité" IDEX2013 (University of Strasbourg), the French Agence Nationale de la Recherche (ANR) under Grant ANR-18-E26-0018 (Project GRICRIS), the Italian Ministry of Universities and Research under Grant PRIN 2022 n. 20229LRAHK (funded by the European Union - Next Generation EU, Mission 4 Component 1, CUP B53D23012680006), the Sapienza University under Ateneo Grants 2023 for project n. RG123188B4CEE028, and the D34Health through the PNC-Spoke3 program research project "Wearable technologies, sensors, and biomarkers for care through digital twin approaches", under Grant B53C22006120001. P. Battigalli gratefully acknowledges the financial support by the ERC (Grant 324219). R. Nagel gratefully acknowledges financial support by FEDER/Ministerio de Ciencia de Innovaci'on (Agencia Estatal de Investigaci'on) through Grant ECO2008-01768, ECO2011-25295, PID2021-125538NB-I00 and through the Severo Ochoa Programme for Centers of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S).

Extended author information available on the last page of the article

Published online: 07 April 2025

JEL Classification C72 · C91

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 Battigalli and Dufwenberg 2022 and the references therein). Beliefs may affect motivation in more than one way. First, as argued by 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.[1]

As documented by Azar (2019), the two most prominent forms of belief-dependent motivation in the theoretical and experimental literature are guilt aversion (Battigalli and Dufwenberg 2007) and intention-based reciprocity (Rabin 1993; Dufwenberg and Kirchsteiger 2004). Here, we offer an innovative theoretical and experimental analysis of these motivations in a game form where they are plausibly salient,[2]: the Trust Game, a stylized social dilemma whereby agent *A* (the truster, "she") takes a costly action that generates a social return, and agent *B* (the trustee, "he") decides how to distribute the proceeds between himself and *A* (Berg et al. 1995; Buskens and Raub 2013). Specifically, we focus on the **Trust Minigame**—a binary version of the Trust Game—in which *A* and *B* are partners in a project with an endowment of €2. Player *A* has to decide whether to *Dissolve* or to *Continue* with the partnership. If player *A* decides to *Dissolve* the partnership, the players split the profit fifty-fifty. If player *A* decides to *Continue* with the partnership, total profit doubles (€4); however, in that case, player *B* has the right to share equally or take everything. In the simultaneous-move game form of Table 1 (the strategic form of the sequential Trust Minigame), player *B* has to state if he would (entirely) *Take* or (equally) *Share* the higher profits before knowing player *A*'s choice; hence, also in the case where *A* chooses *Dissolve*.

Given this game form, we show how *B*-subjects' preferences over distributions of monetary payoffs depend on their second-order beliefs, and how the disclosure of such belief-dependent preferences affects strategic behavior. Differently from the extant literature, we allow subjects' preferences to be simultaneously affected by guilt aversion and reciprocity, and we theoretically analyze behavior both with complete and incomplete information, obtaining most of the testable predictions with an appropriate version of rationalizability for psychological games. Our empirical strategy is to propose a structured questionnaire to elicit (bivariate) psychological types experimentally, and then test the type-dependent theoretical predictions both under incomplete

[1] See, e.g., Berkowitz and Harmon-Jones (2004) and the description of action tendencies in Battigalli et al. (2019). The intellectual home and mathematical framework for models of interacting agents with belief-dependent motivations is an extension of traditional game theory, put forward and labeled "psychological game theory" by Geanakoplos et al. (1989) and further developed by Battigalli and Dufwenberg (2009).

[2] We call "game form" the mere description of the rules of the game, with no specification of players' personal features, such as their preferences.

Table 1 Payoff matrix for the Trust Minigame

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

information and—through disclosure of the questionnaire—in a situation that approximates complete information. Our study addresses the following research questions: Are belief-dependent preferences heterogeneous? Are individual subjects playing the Trust Game better described by the guilt-aversion or the reciprocity model? Is it possible to credibly disclose *B*'s belief-dependent preferences, and do *B*-subjects behave as predicted given their elicited preferences? Does disclosure have the predicted impact on the behavior of matched subjects? In the rest of this introduction we explain our methodology and reseach questions in more detail.

Guilt aversion, reciprocity, and solution concept The first two research questions relate to the simultaneous presence of both belief-dependent motivations. Given their salience, this is an important innovative feature of our paper. On the one hand, **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. On the other hand, according to **intention-based reciprocity**, *B*'s willingness to share depends on his perception of *A*'s costly action as either kind or neutral toward him: The less *A* expects *B* to share, the kinder is her action; therefore, *B*'s willingness to share is *decreasing* in his second-order belief.

Experimental studies of the Trust Game find a positive correlation between elicited second-order beliefs and sharing, consistently with the hypothesis that, in this social dilemma, guilt aversion is an important psychological motivation of *B*-subjects (e.g., Charness and Dufwenberg 2006; Chang et al. 2011, and the studies surveyed in Attanasi and Nagel 2008 and Cartwright 2019).[3] Other experimental studies find evidence consistent with intention-based reciprocity both in the Trust Game (Bacharach et al. 2007; Stanca et al. 2009; Toussaert 2017; Gómez-Miñambres et al. 2021; Rimbaud and Soldà 2024) and in other two-player social dilemmas (e.g., Falk et al. 2008; Dhaene and Bouckaert 2010; Dufwenberg et al. 2011, 2013; Chao 2018; Orhun 2018). Thus, the experimental evidence suggests that both motivations are present in social dilemmas, and especially in the role of trustee in a Trust Game.

The first innovation of the paper is to provide an original theoretical analysis that simultaneously allows for both belief-dependent motivations, and to obtain rationalizability and equilibrium predictions under complete and incomplete information.[4]

[3] See also Guerra and Zizzo (2004), Bacharach et al. (2007), Charness and Dufwenberg (2011), Bracht and Regner (2013), Ederer and Stremitzer (2017), Engler et al. (2018), Attanasi et al. (2019). Experimental studies of other two-player social dilemmas (Dufwenberg and Gneezy 2000; Reuben et al. 2009; Bellemare et al. 2011; Khalmetski et al. 2015; Khalmetski 2016; Di Bartolomeo et al. 2019; Peeters and Vorsatz 2021; Attanasi et al. 2023), and experimental studies of the dictator game (Balafoutas and Fornwager 2017; Morell 2019, and Danilov et al. 2021) also provide support for guilt aversion.

[4] In a game with **complete information** there is common knowledge of (i) the rules of the game, which include how each player is paid as a function of all players' actions, and (ii) players' preferences. If at least one of these conditions fails, there is **incomplete information**.

We make the simplifying assumption that the truster, A, is commonly known to be self-interested, while the trustee, B, has belief-dependent preferences given by a combination of guilt aversion and intention-based reciprocity. Subjects playing in role B had to answer a structured questionnaire. In the main treatment, the answers are made common knolwedge within each matched pair. We assume that this approximates a game with complete information (we expand below on the information regime and its relation with questionnaire disclosure). We rely on our model both to infer belief-dependent preferences from the filled-in questionnaire, and to use such elicited preferences to derive predictions in the Trust Minigame for the complete-information regime (main treatment) and for the incomplete-information regime (control). Since our subjects cannot learn from experience to play an equilibrium, we first look at the implications of the appropriate version of rationalizability (Battigalli et al. 2019, 2020), which in this case obtain in just a few steps. Roughly, as A is commonly known to be selfish, in both regimes the trusting action signals a high belief that B is going to share. With this, if B is highly guilt-averse, he wants to meet A's trust; if instead B is sufficiently close to being selfish, he wants to grab all the surplus. This holds independently of the information regime. However, under complete information A knows whether one of these two cases applies, correctly predicts B's strategy, and acts accordingly. Thus, common knowledge of B's type yields a correlation between rationalizable actions and beliefs of the two players, whereas under incomplete information A's choice and belief are independent of B's psychological type. For intermediate types, B's strategy depends on the precise value of his second-order belief, which rationalizability does not pin down.

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

Information regime and questionnaire disclosure The high heterogeneity of behavior and beliefs found in most experiments on other-regarding preferences (see Cooper and Kagel 2016), especially when these preferences are belief-dependent (see Attanasi and Nagel 2008; Cartwright 2019), makes the assumption that such preferences are common knowledge farfetched. Therefore, although the game rules are made common knowledge in an experiment, it should be assumed that the game subjects play in the laboratory features incomplete information (Attanasi et al. 2016; Battigalli et al. 2019).

A second key feature of our work is the empirical strategy. As in the twin project by Attanasi et al. (2019), we elicit the trustee's belief-dependent preferences through

a structured questionnaire.[5] In the main treatment, the filled-in questionnaire is disclosed and made common knowledge within the matched pair, whereas in the control treatment, the filled-in questionnaire is not disclosed to the truster. The experimental design is such that *B*-subjects should not perceive an incentive to misrepresent their preferences, and indeed we find no significant difference in the pattern of answers across treatments. This supports our main auxiliary assumption: In the treatment with disclosure, *B*'s psychological type is truthfully revealed and made common knowledge; therefore, assuming that *A* is commonly known to be self-interested, this treatment implements a psychological game with complete information.

In this regard, eliciting and disclosing subjects' preferences through a structured questionnaire like the one used in this paper has several methodological advantages. First, belief-dependent preferences are measured and disclosed without data on beliefs. Second, more than one type of belief-dependent preference (e.g., guilt or reciprocity) and their combination are measured and disclosed. Specifically, we estimate bivariate psychological types minimizing the distance between the answers to the hypothetical payback scheme and those resulting from our model according to a parametrized "payback function." Given that this function is not linear, we use a non-linear least square estimation where, to account for the small size of the sample, standard deviations are given by a non-parametric bootstrap technique. The adoption of a bootstrap estimation which relies on a fine grid of values approximating a continuum, constitutes a third methodological advantage of our technique, and is a further innovation of our paper. It improves on other methods that only assess the prevailing belief-dependent preference (Bellemare and Sebald 2023) or deliver only ordinal measures of it (Regner and Harth 2014; Khalmetski et al. 2015; Bellemare et al. 2017, 2018). Finally, the method is easily portable to all asymmetric two-player, or three-player experimental games containing a dictator game as a subgame (see Attanasi et al. 2023). In all such interactions, this experimental technique can be adapted to test complete *versus* incomplete-information predictions of theoretical models with belief-dependent preferences: by comparing behavior in the disclosure *versus* control treatment, one can assess how much of the above-mentioned heterogeneity of behavior and beliefs is only due to diverse belief-dependent preferences or also to the lack of common knowledge about them.

Elicited types and predictions A third key feature of our work lies in the aforementioned combination of the theoretical analysis of the payback scheme with the answer to the questionnaire to estimate bivariate psychological types, in order both to retrieve the distribution of psychological types and to test the type-dependent predictions of the theory. For what concerns the distribution of types, we find that, while preferences are heterogeneous, guilt aversion is indeed the prevalent psychological motivation. For what concerns the test of the theory, we find that behavior and elicited beliefs

[5] We explain the differences with Attanasi et al. (2019) below. Bellemare et al. (2017, 2018) and Khalmetski et al. (2015) elicit the dictator's belief-dependent preferences in a dictator game through a structured questionnaire similar to ours. Regner and Harth (2014) let subjects answer a non-structured post-experimental

Footnote 5 continued

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.

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

As mentioned above, this paper adopts the same questionnaire disclosure technique of Attanasi et al. (2019). The two papers, however, address different questions with different methodologies. Attanasi et al. (2019) focuses on reputation building in a (finitely) Repeated Trust Game. Due to the complexity of the game, (i) theoretically, that article focuses only on guilt aversion and relies on a sequential equilibrium analysis; (ii) empirically, it only allows for mutually exclusive belief-dependent motivations. In this paper, instead, subjects play a one-shot version of the Trust Game. This allows us to *jointly* investigate the effects of guilt and reciprocity, with theoretical predictions mostly based on rationalizability and point estimates of subjects' (bivariate) psychological types, whereas Attanasi et al. (2019) simply classify subjects in different categories according to their payback pattern.

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

2 Design of the experiment

The experimental design is made of three phases and three treatments, summarized in Table 2 (for the experimental instructions see *Online Appendix A.1*). In the first phase, subjects play the Trust Minigame of Table 1, after reporting their beliefs. In the second phase, subjects may be asked to fill in a questionnaire. Then, in the third phase they once again report their beliefs and play the Trust Minigame of Table 1.

Treatments differ only in phase 2 depending on whether (i) subjects playing in role *B* are asked to fill in a questionnaire, and (ii) such answers are disclosed within a rematched *A*-*B* pair. We refer to the treatments, explained 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).[6] The treatment differences of phase 2 influence the information that subjects have when playing the Trust Minigame in phase 3. Indeed, treatment *QD* is assumed to approximate, in phase 3, a situation of complete information on the belief-dependent preferences of *B*-subjects, while both treatments *NoQ* and *QnoD*, and the first phase of *QD*, yield a condition of incomplete information on the belief-dependent preferences of *B* -subjects. Note that we include both types of incomplete-information treatments because we want to control for the possibility that merely filling in the questionnaire affects subjects' choices.

[6] Further details of experimental procedures can be found in the *Online Appendix A.2*.

Table 2 Summary of the Experimental Design

	Treatments		
	NoQ (40 paris)	*QnoD* (40 paris)	*QD* (80 paris)
Phase 1		Trust Minigame with Beliefs Eliciation	
Phase 2	**No Q**uestionnaire	**Q**uestionnaire with **no D**isclosure	**Q**uestionnaire with **D**isclosure
Phase 3		Trust Minigame with Beliefs Eliciation	
		Final Questionnnaire with *no D*isclosure	

Table 3 Questionnaire (Hypothetical Payback Scheme) in phase 2

A thought you would have chosen *Share* with probability	Your payback (in €)
0%	Between 0.00 and 4.00
10%	Between 0.00 and 4.00
…	…
90%	Between 0.00 and 4.00
100%	Between 0.00 and 4.00

At the beginning of every experimental session, each of the 20 participants, or subjects, is randomly assigned with equal probability to role *A* (*A*-subject) or role *B* (*B*-subject) of the Trust Minigame. This determines 10 *A*-*B* pairs in each session. Each subject maintains the same role until the end of the session. Participants are told that the experiment is made of three phases. Instructions of each new phase are given and read aloud only prior to that phase. We now describe the three phases in detail.

Phase 1 This phase is the same for all treatments and consists of a random matching between *A*-subjects and *B*-subjects, and two subsequent decision tasks:
Elicitation of beliefs With regard to the Trust Minigame of Table 1: Each *A*-subject is asked to guess the percentage of *B*-subjects in her session who will choose *Share* (*A*'s *initial first-order belief*). Each *B*-subject is asked to guess the answer of his co-paired *A* about the percentage of *B*-subjects who will choose *Share* (a feature of *B*'s *unconditional second-order belief*), and to guess the choice—*Dissolve* or *Continue*—of the co-player (a feature of *B*'s *first-order belief*).
Choice Within each pair, player *A* and player *B* simultaneously make their choice in the Trust Minigame of Table 1.

At the end of phase 1, subjects receive no information feedback on the two decision tasks. Indeed, at the beginning of phase 1, they were informed that the gains in the belief-elicitation task and in the Trust Minigame would be communicated at the end of the experiment.

Phase 2 In *NoQ*, subjects proceed directly to phase 3. In *QnoD* and *QD*, subjects are randomly re-matched to form other 10 pairs (absolute-stranger matching design). Each *B*-subject is asked to fill in the questionnaire of Table 3 (*hypothetical payback scheme*), considering the following hypothetical situation: His new *A*-co-player has chosen *Continue* and he, *B*, has chosen *Take*, thereby earning €4 and leaving *A* with

€0. Given this, *B* has the possibility—if he wishes—to give part of this amount back to *A*. He is allowed to condition his payback on the hypothesized first-order belief of *A*.

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

B-subjects first fill in the questionnaire on paper and then have to copy the answers on their computers. *A*-subjects read and listen to the instructions of phase 2. Among the subjects of each *QnoD* and *QD* session, it is made public information that neither the responding *B*-subject nor anyone else will receive any payment for the answers to the questionnaire. Furthermore, in *QnoD* it is public information that *B*'s filled-in questionnaire *will not be* disclosed to anyone. On the other hand, in *QD* it is public information that *B*'s filled-in questionnaire *will be* disclosed to a randomly-chosen *A*-subject. Actually, this subject is the one randomly matched with *B* at the beginning of phase 2. At the end of this phase, the matched *B*'s filled-in questionnaire appears on *A*'s screen, and the latter is invited to copy it on paper. At this stage, subjects do not know yet that in phase 3 they are going to play again the Trust Minigame, with the same match of phase 2.

Phase 3 Also this phase is the same for all treatments and it consists of the same two decision tasks of phase 1, with a new random matching. Specifically, in *NoQ* subjects are randomly re-matched to form other 10 pairs; in *QnoD* and *QD*, each *A*-subject is matched with the same *B*-subject as in phase 2. In *QnoD* and *QD*, each *B*-subject can keep his previously filled-in paper questionnaire with him for the duration of this phase. In *QD*, *A* can keep the matched *B*'s filled-in questionnaire (previously copied on paper) with her. At the beginning of phase 3 of *QD*, it is made public information that, in each pair, *B*'s filled-in questionnaire disclosed at the end of phase 2 corresponds to the matched *B*-subject of phase 3.[7] At the end of phase 3, in *QD* and *QnoD* all filled-in questionnaires are collected by the experimenter.

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

Payment Results of both phase 1 and phase 3 are communicated after the final questionnaire. Each subject learns the co-player's choice in the Trust Minigame in phase 1 and in phase 3, and whether her first-order belief (*A*-subject) or his first and second-order beliefs (*B*-subject) in phase 1 and in phase 3 are correct. Subjects are paid for choices and beliefs in phase 1 and in phase 3. Specifically, for each phase, choices are paid according to the payoff matrix of Table 1. In addition, (i) subjects in role *A* receive 5 € if their guess of the percentage of *B* subjects who choose *Share* is

[7] On such "unexpected data use," see Charness et al. (2022), and, in particular, Sect. 5.

correct, and (ii) subjects in role B receive 5€ if both their conjecture on A's guess, and their conjecture on A's choice are correct.

3 Model

In this section, we put forward a portable model of belief-dependent preferences with guilt aversion and intention-based reciprocity (Sect. 3.1). Then we use it to derive predictions for the Trust Minigame (Sect. 3.2), both under complete information (Sect. 3.2.2) and incomplete information (Sect. 3.2.3). Finally, we present the testable predictions (Sect. 3.3).

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 and Dufwenberg (2007, 2009), we allow a player's preferences over outcomes to depend on the beliefs of the co-player, which yields a simpler representation. Higher-order beliefs appear in the expected utility-maximization problems embedded in solution concepts.

Specifically, we represent a player's preferences with a psychological utility function that depends only on (m_i, m_j) and on the co-player's first-order beliefs (which include the co-player's plan of action, a belief about what he/she is going to do). Let α_j denote j's first-order belief about behavior. The latter is represented by a strategy pair (s_j, s_i), and the marginal of α_j on S_j represents j's plan. We obtain a utility function of the form $u_i(m_i, m_j, \alpha_j)$ by assuming that i dislikes disappointing j (the "guilt" component), and cares about the monetary payoff distribution that j expects to achieve (the "intention-based reciprocity" component); both variables depend on α_j.

We maintain the assumption that players have deterministic plans.[8] With this, let s_j be the plan (pure strategy) of player j, then α_j is determined by the pair (s_j, α_{ji}), where α_{ji} is j's belief about i's behavior, 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 $\mathbb{E}_A[\widetilde{m}_A; \alpha_A] = 2 \times 0.6 = 1.2$.[9] The psychological utility of B depends on this expectation. Of course, since B does not know α_A, his valuation of (m_B, m_A) is the subjective expectation $\mathbb{E}_B[u_B(m_B, m_A, \widetilde{\alpha}_A)]$ according to his second-order belief. Next we provide the details of our specification of the psychological utility function $u_B(m_B, m_A, \alpha_A)$.

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

[8] Note that subjective expected utility maximizing players have no strict incentive to randomize.

[9] We use a tilde over a math symbol to denote a random variable. For example, since A does not know the payoff she is going to get, this number is a random variable from her point of view, and its expectation is $\mathbb{E}_A[\widetilde{m}_A; \alpha_A]$.

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}: $K_j(\alpha_j) = \mathbb{E}_j[\widetilde{m}_i; \alpha_j] - m_i^e(\alpha_{ji})$.

We jointly analyze these two belief-dependent motivations, assuming that i's **psychological utility function** has the following additively separable form:

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

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

Next we move to the specific analysis of the Trust Minigame. We assume that preferences are role-dependent. In particular, A (the truster) has *selfish* risk-neutral preferences, i.e., eq. (1) reduces to $u_A = m_A$. As for B, we assume that his utility may display both guilt aversion and reciprocity. The assumption that the belief-dependent component of the utility function is activated by the context (e.g., by playing in a particular role) is consistent with the evolutionary psychology of emotions (e.g., Haselton and Ketelaar 2006), which highlights how contextual cues moderate the effects of emotions, and also with the conceptual act theory of emotion (e.g., Barrett 2006). A more extensive discussion of the psychological and physiological evidence in favor of the role-dependent assumption can be found in Attanasi et al. (2016). Additionally, note that Bicchieri et al. (2011) show that trustworthiness is a social norm, while trusting behavior is not. This is further evidence that drivers of trust and trustworthiness are different, and that it is plausible to assume role-dependent preferences.[10]

We rely on belief-dependent preferences not only to analyze binary allocation choices (as, e.g., in Attanasi et al. 2016), but also to analyze the hypothetical payback scheme shown in Table 3 above. As discussed in Attanasi et al. (2019), when considered separately, guilt aversion and reciprocity have opposite effects on the payback scheme. Guilt aversion implies that the payback function is increasing in $\alpha_{AB}(Share)$, as the more A expects B to *Share*, the higher her expected payoff, and therefore her disappointment when she receives less. Reciprocity, on the contrary, implies that the payback function is decreasing in $\alpha_{AB}(Share)$, as A's choice to *Continue* is expected to give more to B, hence is kinder, when $\alpha_{AB}(Share)$ is lower. The combination of guilt and reciprocity may also yield a U-shaped payback function. The analysis of the payback scheme is contained in *Online Appendix B.1*.

Differently from Attanasi et al. (2019), this paper analyzes, both theoretically and experimentally, the interplay between guilt aversion and reciprocity. For this reason and for the sake of tractability, in both choices that B-subjects are asked to make—in

[10] If we allow A to be guilt-averse, a new type of behavior may arise in which A *Continues* despite assigning a very low probability to *Share*, not to disappoint B, if B is believed to assign a high probability to *Continue* (see Attanasi et al. 2016).

the Trust Minigame and in the hypothetical payback scheme—, we use a parametric specification of eq. (1) with the following features:

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

To sum up, we assume the following functional form:

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

where G_i and R_i respectively parametrize sensitivity to guilt and reciprocity.[11] This parametrization achieves a good balance between tractability and flexibility (see *Online Appendix B.1*).

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*. This implies that the only relevant uncertainty for *B* (conditional on *Continue*) is the initial belief of *A* about *B*'s strategy, $\alpha_{AB}(Share)$.

3.2 Theoretical predictions for the Trust Minigame

Since we assume that *B* chooses as if he had observed the trusting action *Continue*—the only situation in which *B*'s choice matters—, we apply solution concepts for the sequential Trust Minigame where *A* moves first and *B* observes *A*'s choice. We consider two situations: the complete-information regime of common knowledge of the psychological utility function u_B in (2), which we approximate in the lab in phase 3 of the main treatment (*QD*), and the incomplete-information regime where u_B is not common knowledge, which is the standard situation in experiments. On top of assuming that *A* is selfish and risk-neutral, we also assume that this is common knowledge. Given this, the behavioral implications of rationality and strong belief in rationality (given by the first two steps of rationalizability) are common to the complete- and

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

incomplete-information regimes. These results are presented in Sect. 3.2.1 and summarized in Fig. 1a.

Then, we focus on each information regime separately. In Sect. 3.2.2, we first provide a rationalizability analysis of the complete-information regime based on forward-induction reasoning (cf. Sect. 5 of Battigalli and Dufwenberg 2009, and Battigalli et al. 2020). Since rationalizability does not yield sharp predictions for all possible cases (parameters of u_B), we also provide refined predictions based on equilibrium analysis (see Fig. 1b). In Sect. 3.2.3, we turn to incomplete information. Here, the behavioral predictions of rationalizability turn out to coincide with the first two steps of Sect. 3.2.1, and are therefore weaker than under complete information. We refine these predictions, to some extent, by considering Bayesian Nash equilibria.

3.2.1 Rationalizability with forward induction: the first two steps

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

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

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

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

Rationality of A Since we are analyzing a psychological game where the utility function of B depends on the first-order beliefs of A, we use a notion of rationalizability that gives (either partial or sharp) predictions about the behavior, s_A, and first-order belief of A.[12] Writing $\alpha = \alpha_{AB}(Share)$ to ease notation, the set of behavior-belief pairs consistent with A's rationality (assumption 1 above) is

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

Rationality of B As for B, we have to consider his **psychological type** (G, R) (the parameter vector that identifies u_B) and define the set of triples $(s_B; G, R)$ consistent

[12] A second reason to give predictions about $(s_A, \alpha_{AB}(Share))$ is that we elicit $\alpha_{AB}(Share)$, which is therefore "observable."

with assumptions 1 and 2 above. We consider predictions about $(s_B; G, R)$ because, if A thinks strategically, she forms beliefs about how s_B is related to (G, R).[13] In the Trust Minigame, $D_A(\alpha_A, m_A) = \max\{0, 2\alpha - m_A\}$ and $K_A(\alpha_A) = \frac{3}{2} - \alpha$. Plugging these disappointment and kindness functions in (2), we obtain

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

where, conditional on *Continue*, $(m_A, m_B) = (2, 2)$ if player B chooses *Share* and $(m_A, m_B) = (0, 4)$ if he chooses *Take*. Therefore, player B chooses *Share* if and only if $\mathbb{E}_B[u_B(2, 2, \widetilde{\alpha})|Cont] \geq \mathbb{E}_B[u_B(4, 0, \widetilde{\alpha})|Cont]$ according to eq. (3), that is, writing $\beta = \mathbb{E}_B(\tilde{\alpha}|Cont)$,

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

With this, we note that, w.l.o.g., *we can analyze the "willingness-to-share" of B as if he were certain of A's first-order belief* α conditional on observing *Continue*. Therefore, in the analysis of rationalizability we reason as if B had a point belief $\beta \in [0, 1]$ about α conditional on *Continue* (thus, here the meaning of symbol β is a special case of the conditional expectation $\mathbb{E}_B[\widetilde{\alpha}|Cont]$). With this, inequality (4) becomes

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

Our analysis depends on the shape of B's **willingness-to-share** function $WS(\beta; G, R)$ implied by psychological type (G, R).[14] Clearly, *Share* is justifiable as a best reply for B of type (G, R) if $WS(\beta; G, R) \geq 0$ for some $\beta \in [0, 1]$, that is, if $\max_{\beta\in[0,1]} WS(\beta; G, R) \geq 0$; similarly, *Take* is justifiable for B of type (G, R) if $\min_{\beta\in[0,1]} WS(\beta; G, R) \leq 0$. Conversely, if $\min_{\beta\in[0,1]} WS(\beta; G, R) > 0$ then *Share* is the only justifiable choice, that is, the dominant choice for (G, R); if instead $\max_{\beta\in[0,1]} WS(\beta; G, R) < 0$ then *Take* is the dominant choice for (G, R). Rationality implies that player B of type (G, R) chooses the dominant action when it exists. This gives the step-1 prediction set P_B^1.

Forward induction First, note that A's choice *Continue* is consistent with A's (selfish) rationality, because A may subjectively believe that *Share* is more likely than *Take*.[15] Therefore the assumption that B strongly believes in A's rationality implies that B is certain that $\alpha \geq 1/2$ conditional on *Continue*; formally,

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

[13] See, e.g., how rationalizability is defined in Battigalli et al. (2020). Furthermore, we experimentally identify (G, R). Hence we can test these joint predictions.

[14] See Fig. B.2 in *Online Appendix B.2*.

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

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

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

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

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

The foregoing analysis leads to a related question: When is it the case that, for B of type (G, R) who strongly believes in A's rationality, *Share* (respectively, *Take*) is the unique best reply independently of the specific belief of B? In other words, when is a strategy of B "forward-induction (FI) dominant" for psychological type (G, R)? The answer is that *Share* (respectively, *Take)* is FI-dominant for (G, R) if and only if $WS(\beta; G, R) > 0$ (respectively $WS(\beta; G, R) < 0$) *for every* $\beta \geq 1/2$, which is equivalent to $\min_{\beta \in \left[\frac{1}{2}, 1\right]} WS(\beta; G, R) > 0$ (respectively, $\max_{\beta \in \left[\frac{1}{2}, 1\right]} WS(\beta; G, R) < 0$). Thus, we obtain the following **FI-dominance regions** in the space of psychological types (G, R), represented in Fig. 1a[16]

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

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

Finally, by definition, $\{Share\} \times \mathbb{S} \subseteq P_B^{2,S}$ and $\{Take\} \times \mathbb{T} \subseteq P_B^{2,T}$.

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

3.2.2 Complete information

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

Rationalizability Under complete information, the psychological type (G, R) of B is common knowledge. Therefore, rationalizability yields sharp predictions when (G, R) belongs to an FI-dominance region of Fig. 1a. If A believes in B's rationality and B's strong belief in rationality (assumptions 1 and 2 above), and $(G, R) \in \mathbb{S}$, then

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

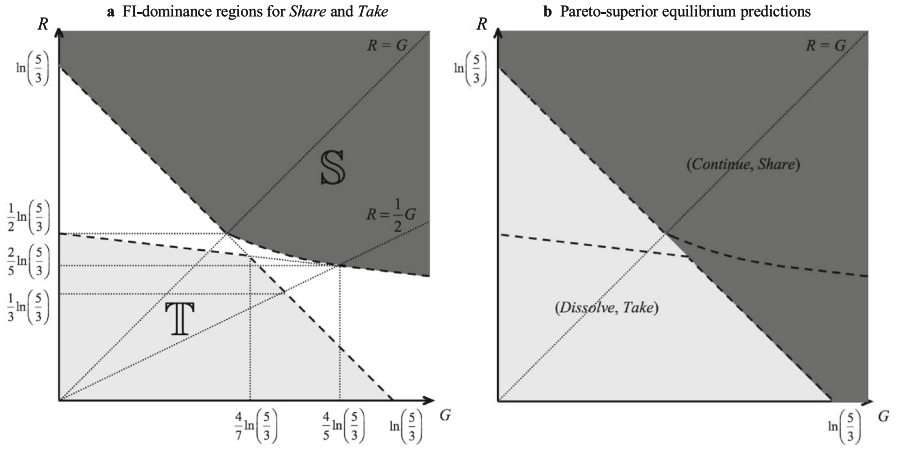

Fig. 1 Theoretical predictions

A is certain of *Share* ($\alpha = 1$) and plays *Continue*; if $(G, R) \in \mathbb{T}$, then A is certain of *Take* ($\alpha = 0$) and plays *Dissolve*. If B anticipates this and $(G, R) \in \mathbb{S}$, then he is initially certain of *Continue* and that $\alpha = 1$, and he is also certain that $\alpha = 1$ conditional on *Continue*. If instead $(G, R) \in \mathbb{T}$, then B is initially certain of *Dissolve* and that $\alpha = 0$, but strong belief in rationality implies that he would be certain that $\alpha \geq 1/2$ if he—unexpectedly—observed *Continue*.

Conversely, when the psychological type of B does not belong to any FI-dominance region (i.e., it is in the white intermediate region of Fig. 1a), then rationalizability does not yield predictions about behavior: even if B strongly believes in A's rationality and therefore is certain that $\alpha \geq 1/2$ conditional on *Continue*, for each $(G, R) \notin \mathbb{S} \cup \mathbb{T}$ there is some $\beta \geq 1/2$ that makes *Share* a best reply and also some (other) $\beta \geq 1/2$ that makes *Take* a best reply. Thus, the behavior of B and the belief and behavior of A are not pinned down. The following proposition summarizes the behavioral predictions of rationalizability. Since in our experiment we do not measure the conditional second-order beliefs of B-subjects, we focus on predictions about (s_A, s_B, α). When such predictions are sharp, then $\alpha \in \{0, 1\}$ and the unconditional (i.e., initial) second-order belief of B coincides with α.

Proposition 1 *Under complete information, the prediction of rationalizability based on forward induction is as follows:*

(i) Continue, Share, *and* $\alpha = 1$ *if* $(G, R) \in \mathbb{S}$,
(ii) Dissolve, Take, *and* $\alpha = 0$ *if* $(G, R) \in \mathbb{T}$,
(iii) *any* (s_A, s_B, α) *such that* s_A *is a best reply to* α *(i.e.,* $(s_A, \alpha) \in P_A^1$*) is possible if* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$.

Equilibrium analysis To sharpen our predictions we turn to equilibrium analysis. Since we assume that B chooses as if he had observed *Continue*, we analyze the Perfect Bayesian Equilibria (PBE) of the sequential Trust Minigame with complete information (cf. Battigalli and Dufwenberg 2009). In a PBE initial beliefs are correct,

A best responds to her initial first-order belief α, and B best responds to his conditional (second-order) belief about α, which coincides with the unconditional second-order belief when *Continue* has positive probability.[17] Mixed or partially mixed equilibria are often justified as stable states of learning dynamics, but such justification is precluded here because we consider one-shot interactions. Pure equilibria can instead be justified (sometimes) as outcomes of strategic reasoning. Therefore, we focus on pure PBE's.

We begin with a preliminary observation. If the psychological type of B belongs to the region of the parameter space of Fig. 1a where *Share* is dominant, then backward induction implies that the only PBE strategy of B is *Share*; hence, the unique PBE is the "trust equilibrium" $(Continue, Share, \alpha = \beta = 1)$. Similarly, if the psychological type of B belongs to the region where *Take* is dominant, then the only PBE strategy of B is *Take*; hence, all PBE's are of the "no-trust" kind with $\alpha = 0$ and $(s_A, s_B) = (Dissolve, Take)$.[18] Thus, for all the aforementioned types of B, the PBE prediction is unique and coincides with the complete-information rationalizability prediction.

Now suppose that $WS\,(1; G, R) > 0$, but $WS\,(\beta; G, R) \leq 0$ for *some* $\beta < 1$. Then, there are multiple pure-strategy PBE's, the "trust equilibrium" and the "no-trust equilibria" mentioned above.[19] In particular, "no-trust" is an equilibrium for each (G, R) outside the FI-dominance region $\mathbb{S}$: by definition, if $(G, R) \notin \mathbb{S}$ there is some $\beta \geq 1/2$ such that $WS\,(\beta; G, R) \leq 0$; hence, B is willing to *Take* even if he rationalizes the possibly unexpected choice *Continue*, which implies that there is a pure PBE of the form $(Dissolve, Take, \alpha = 0, \beta \geq 1/2)$ satisfying forward induction.

To obtain sharp predictions, in the case of multiplicity we apply a *Pareto-selection criterion:*[20] we assume that the pure equilibrium with higher payoffs for both players is salient and therefore players' expectations are coordinated on such equilibrium. We show that this is consistent with our complete-information rationalizability analysis based on forward-induction reasoning; hence, we are indeed refining the rationalizability predictions.

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

Proposition 2 *The Pareto-superior, pure equilibrium prediction under complete information is as follows:*

(i) Continue, Share, *and* $\alpha = 1$ *if* $G + R \geq \ln(5/3)$,

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

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

[19] Psychological games have multiple PBE's even in situations where standard games have a unique PBE; the Trust Minigame is a case in point (Geanakoplos et al. 1989; Battigalli and Dufwenberg 2007, 2009).

[20] In this psychological game, higher equilibrium material payoffs imply higher equilibrium psychological utilities. Della Lena et al. (2023) use the same selection criterion.

(ii) Dissolve, Take, *and* $\alpha = 0$ *if* $G + R < \ln(5/3)$.
These predictions refine the complete-information rationalizability predictions based on forward induction.

Figure 1b represents the regions of the space of psychological types (G, R) with the Pareto-superior equilibrium prediction of (*Continue*, *Share*) and (*Dissolve*, *Take*) according to Proposition 2. Note that the locus of $G + R = \ln(5/3)$ is a line that separates the FI-dominance regions $\mathbb{S}$ and $\mathbb{T}$ in Fig. 1a.

3.2.3 Incomplete information

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

Rationalizability We use a rationalizability concept for games with partially unknown utility functions, which characterizes the implications of rationality and common strong belief in rationality.[21] Steps 1 and 2 for player B are already given in Sect. 3.2.1: the set of possible triples $(s_B; G, R)$ consistent with rationality and strong belief in rationality is $P_B^2 = P_B^{2,S} \cup P_B^{2,T}$. Furthermore, if $(G, R) \in \mathbb{S}$ then B certainly chooses *Share*, and if $(G, R) \in \mathbb{T}$ then B certainly chooses *Take,* whereas if (G, R) does not belong to either FI-dominance region then both strategies can be justified by a conditional second-order belief consistent with the assumption that A is rational.

Since we are not positing any specific assumption concerning A's exogenous (ex ante) beliefs about the parameter vector (G, R), we cannot derive any further implication about A's behavior. To see this, note that if A assigns more than 50% probability to $\mathbb{S}$, then $\alpha > 1/2$ and the best reply is *Continue*, if instead A assigns more than 50% probability to $\mathbb{T}$, then $\alpha < 1/2$ and the best reply is *Dissolve.* Since step 3 does not refine the predictions for A, the incomplete-information rationalizability algorithm stops, i.e., it gives the same predictions at each further step for each player. The following proposition summarizes:

Proposition 3 *Without restrictions on exogenous beliefs, incomplete-information rationalizability implies (only) that* $(s_A, \alpha) \in P_A^1$ *and* $(s_B; G, R) \in P_B^2$*; in particular,* B *chooses* Share *if* $(G, R) \in \mathbb{S}$ *and* Take *if* $(G, R) \in \mathbb{T}$*, while both strategies are rationalizable for* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$.

Equilibrium analysis We first need to introduce some terminology. We call "**exogenous**" an initial (pre-play) belief about an exogenous variable or a parameter. In particular, a belief about (G, R) is an exogenous first-order belief of A. We call "**endogenous**" a belief about a variable that we try to explain with the strategic analysis of the game. Specifically, α is the endogenous first-order belief that determines A's choice, and the cumulative distribution functions $F_B(x) = \mathbb{P}_B(\widetilde{\alpha} \leq x)$, $F_B(x|Cont) = \mathbb{P}_B(\widetilde{\alpha} \leq x|Cont)$ are—respectively—the unconditional and conditional endogenous second-order beliefs of B (cf. Attanasi et al. 2016). Bayesian

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

equilibrium analysis rests on specific assumptions about players' exogenous beliefs (cf. Harsanyi 1967–68). The only behavioral implications of equilibrium analysis that are robust with respect to such assumptions are those given by incomplete-information rationalizability.[22] To refine such predictions with equilibrium analysis we thus have to posit some restrictions on players' exogenous beliefs and convert them to restrictions on the distribution of behavior and endogenous beliefs.

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

We first list and motivate our qualitative assumptions about exogenous beliefs, then, Proposition 4 summarizes our qualitative predictions.

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

(2) *[B-heterogeneity]* It is even more plausible that B-subjects have *heterogeneous and dispersed exogenous second-order beliefs* about the exogenous first-order beliefs of the A-subjects. Furthermore, B-subjects *believe that assumption (1) holds.* Thus, in particular, they believe that a positive fraction of A-subjects *Continue*.

(3.i) *[Independence between roles]* When subjects are matched at random and do not observe anything about the other subject with whom they are matched, the type of A must be *independent* of the type of B.

(3.ii) *[Independence within roles]* Furthermore, we assume that the psychological type of B and his hierarchy of exogenous beliefs are *independent*.[23]

The following proposition summarizes our qualitative predictions:

Proposition 4 *Under the stated assumptions, in every equilibrium of the Trust Minigame with incomplete information a positive fraction of* A*-types choose* Continue; *furthermore:*

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

[23] This assumption of independence between psychological types and hierarchies of exogenous beliefs rules out the possibility of a false-consensus effect (see Ross et al. 1977). Yet the alternative assumption that higher types of B hold (stochastically) higher second-order beliefs is not validated by our data. See statement (3.ii) in Sect. 4.3.

(1) *[A-heterogeneity]* A*-types have heterogeneous, dispersed beliefs* α *about* B*'s strategy, hence, a substantial fraction of* A*-types have* α *well above* 0 *and well below* 1*;*

(2) *[B-heterogeneity]* B*-types have heterogeneous, dispersed initial beliefs about* A*'s strategy and* α*; conditional second-order beliefs are also heterogeneous, but have support in* $[1/2, 1]$.

(3.i) *[Independence between roles] The strategy and beliefs of* A *are independent of the strategy, psychological type, and beliefs of* B*;*

(3.ii) *[Independence within roles]* B*'s first- and second-order beliefs are independent of the psychological type;*

(4) *[FI-dominance]* B*-types with high values of* G or R *(i.e., with* $(G, R) \in \mathbb{S}$*) choose* Share*,* B*-types with low values of* G and R *(i.e., with* $(G, R) \in \mathbb{T}$*) choose* Take;

(5) *[Choice-belief correlation] The choice of intermediate types* t_B *depends on the equilibrium conditional belief* $\mathbb{P}_{t_B}(\cdot|Cont)$*; in particular, the proportion of* B*-types with* $G \geq 2R$ *who choose* Share *is positively correlated with the conditional second-order belief* $F_{t_B}(\cdot|Cont)$.

Results (1)–(3.ii) follow from the corresponding assumptions. To understand the intuition behind result (4), let $\mathbb{P}_{t_i}(\cdot)$ denote the equilibrium beliefs of type t_i of player i. Assumption (2) implies that, for each type t_B, $\mathbb{P}_{t_B}(Cont) > 0$ so that $\mathbb{P}_{t_B}(\cdot|Cont)$ is well defined. In equilibrium, an A-type t_A *Continues* only if $\alpha_{t_A} = \mathbb{P}_{t_A}(Share) \geq 1/2$. Therefore, for each t_B, the equilibrium conditional belief must satisfy the forward-induction requirement $\mathbb{P}_{t_B}(\widetilde{\alpha} \geq 1/2|Cont) = 1$. This in turn implies that the equilibrium predictions coincide with the rationalizability ones of Proposition 3 for psychological types of B in the FI-dominance regions $\mathbb{S}$ and $\mathbb{T}$.

Finally, given assumptions (1)-(2) about the dispersion of exogenous beliefs, one can also show that the distributions of α, $\mathbb{E}_B[\widetilde{\alpha}]$, and $\mathbb{E}_B[\widetilde{\alpha}|Cont]$ are dense in sub-intervals of—respectively—$[0, 1]$, $[0, 1]$, and $[1/2, 1]$, that is, there is a large fraction of subjects with "intermediate" beliefs (taking into account the FI-requirement for $\mathbb{E}_B[\widetilde{\alpha}|Cont]$). The behavior of B-types t_B with psychological type (G, R) out of the FI-dominance regions depends on their equilibrium conditional belief $\mathbb{P}_{t_B}(\cdot|Cont)$. Taking into account that function $WS(\beta; G, R)$ is increasing on $[1/2, 1]$ if and only if $G \geq 2R$ (see eq. (5)), assumption (3.ii) implies that for every psychological type (G, R) with $G \geq 2R$, a higher conditional second-order belief $F_{t_B}(\cdot|Cont)$(in the sense of stochastic dominance) yields a higher willingness to share $\mathbb{E}_{t_B}[WS(\widetilde{\alpha}; G, R)|Cont]$, hence result (5) follows.

3.3 Theoretical predictions and experimental design

The theoretical analysis in Sects. 3.2.2 (complete information) and 3.2.3 (incomplete information) leads to several testable predictions. These predictions are related to B's psychological type, elicited through the questionnaire of phase 2 (final questionnaire for *NoQ*). Answers to the questionnaire are supposed to reveal whether B's preferences

are belief-dependent and whether guilt or reciprocity is the prevailing motivation (see *Online Appendix B*).

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

- *Phase 3 of Treatment QD:* The questionnaire filled in by B is disclosed to the matched A-subject and made common knowledge within the matched pair. Assuming that the filled-in questionnaire identifies B's psychological type and that A is commonly known to be selfish, the matched subjects play a psychological game with *complete information*.
- *Treatments NoQ, QnoD; Phase 1 of Treatment QD*: A obtains no information about B. Therefore the matched subjects play a psychological game with *incomplete information*.

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

1. *Complete information* (*phase 3 of QD*) Under disclosure of the filled-in questionnaire, we predict a polarization of behavior and beliefs because common knowledge of B's psychological type works as a coordination device. If B is sufficiently selfish (low guilt and/or reciprocity parameters, $(G, R) \in \mathbb{T}$), the unique rationalizable prediction is (*Dissolve*, *Take*) and $\alpha = 0$. If B is sufficiently other-regarding (high guilt and/or reciprocity parameters, $(G, R) \in \mathbb{S}$), the unique rationalizable prediction is (*Continue*, *Share*), and $\alpha = 1$ (see Fig. 1a). Such predictions are refined by the Pareto-superior equilibrium (see Fig. 1b), according to which low (respectively, high) trust prevails if $G + R < \ln(5/3)$ (respectively, $G + R \geq \ln(5/3)$). See Propositions 1 and 2.
2. *Incomplete information* (*all other phase-treatment combinations*) Without disclosure, there are more heterogeneity of behavior and more dispersed beliefs. A first cause of this heterogeneity is that, by random matching, under incomplete information behavior and beliefs of A-subjects are independent of behavior and beliefs of B -subjects. As a consequence, we cannot observe the polarization on either (*Dissolve*, *Take*) and $\alpha = 0$, or (*Continue*, *Share*) and $\alpha = 1$, that arises under complete information. A second cause of heterogeneity is the presence of "intermediate" beliefs. This is quite obvious for A-subjects (assuming heterogeneous, dispersed beliefs about B's psychological type). More interestingly, there is a parameter region with intermediate values of G and low values of R ($G + R > \ln(5/3)$, $(G, R) \notin \mathbb{S}$, see Fig. 1a) where B -subjects would cooperate and hold high second-order beliefs under the complete-information Pareto-superior equilibrium (see Fig. 1b), while they exhibit less cooperative behavior and intermediate second-order beliefs under the incomplete-information Bayesian equilibrium (see Propositions 3 and 4). Symmetrically, there is also a parameter region with intermediate values of R and low values of G ($G + R < \ln(5/3)$, $(G, R) \notin \mathbb{T}$, see Fig. 1a) where the opposite happens, i.e., these types may cooperate under incomplete information, but not under the complete-information Pareto-superior equilibrium. We say that "**guilt prevails for FI-underdetermined subjects**" if the fraction of subjects with utility type in the

latter region is small compared to the fraction of subjects with utility type in the former region.

3. *Complete versus incomplete information* Rationalizability yields the same behavioral predictions (or lack thereof) for *B*-subjects under both complete and incomplete information (compare Propositions 1 to 3). Yet, we also consider equilibrium predictions, which differ across information scenarios (compare Proposition 2 to Proposition 4). Therefore, we rely on such predictions to qualitatively compare players' behavior and beliefs across treatments and phases.
   The comparison between complete- and incomplete-information regimes can be made *between subjects*, by comparing phase 3 of *QD versus NoQ* and *QnoD*, and also *within subjects*, by comparing phase 3 *versus* phase 1 of *QD*. Note that we expect no difference between phase 1 and phase 3 of *NoQ* and *QnoD*, as they both yield incomplete-information.
   First, points 1 and 2 above imply that behavior and beliefs are polarized under complete information but not under incomplete information.
   A second comparative prediction concerns the extent of cooperation. This crucially depends on whether guilt prevails among FI-underdetermined subjects. In the region where guilt prevails, high second-order beliefs are associated to high incentives to *Share*, and therefore we expect more cooperation when beliefs are polarized, i.e., in the complete-information regime. When reciprocity prevails, high second-order beliefs are associated with low incentives to *Share*, and we expect less cooperation when beliefs are polarized. As a consequence, the presence of intermediate second-order beliefs in the incomplete-information regime decreases cooperation in the former region and increases it in the latter one. Since there is evidence in the literature that guilt aversion is the most important psychological motivation in Trust-Game contexts,[24] we expect guilt to prevail among FI-underdetermined subjects. Therefore, we predict more cooperative behavior of *B*-subjects under complete-information.

4 Data analysis

Here we present and discuss our experimental data in light of the theoretical model. Relying on the hypothetical payback scheme introduced in Table 3, we first present the categorization of *B*'s belief-dependent preferences derived from the answers to the questionnaire of Table 3 (Sect. 4.1). With this in mind, we analyze *A*'s and *B*'s behavior (including their side bets, hence their elicited beliefs) in the Trust Minigame. We first use the complete-information predictions to analyze subjects' behavior in phase 3 of the treatment with questionnaire disclosure, *QD* (Sect. 4.2). We then use the incomplete-information predictions to analyze behavior in phase 1 of *QD* and in the treatments without questionnaire disclosure, *NoQ* and *QnoD* (Sect. 4.3), and compare behavior in all these phase-treatment combinations with behavior in phase 3 of *QD* (Sect. 4.4). Finally, we discuss an alternative classification of types (Sect. 4.5).

[24] See Bellemare et al. (2017), Attanasi et al. (2019), Cartwright (2019).

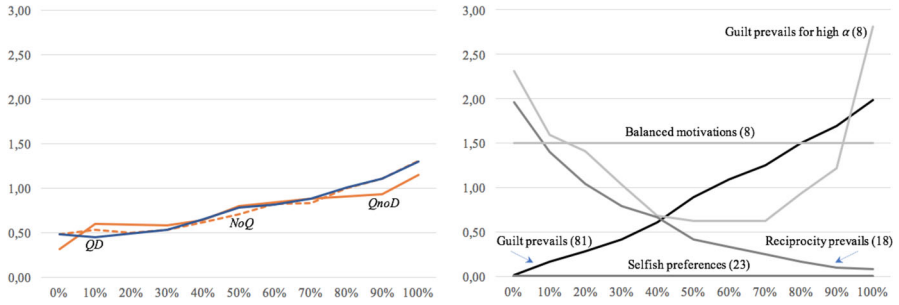

Fig. 2 *B*'s average payback pattern, by treatment (left panel) and type (right panel). The left panel reports *B*'s average payback pattern in *NoQ* (40 subjects), *QnoD* (40 subjects), and *QD* (80 subjects). The right panel reports the average payback pattern of *B*-subjects according to the five possible quasi-convex shapes of $\xi(\alpha)$ predicted by our model of guilt and reciprocity; for each average pattern, the intensity of the black color indicates the relative frequency of the corresponding shape in the population of *B*-subjects (reported in parentheses)

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

The experimental elicitation of *B*'s belief-dependent preferences in the Trust Minigame relies on his answers to the questionnaire of Table 3 (see Sect. 2). We call "**payback pattern**" the actual answers of a *B*-subject, with one payback value for each hypothesized α (*A*'s belief about *B*'s strategy *Share*). Recall that the payback pattern gives 11 observations for *B*'s payback function, i.e., one for each $\alpha \in \{0, 10\%, \ldots, 100\%\}$. In *Online Appendix C* we report payback patterns of the 160 *B*-subjects in our experiment.

The left panel of Fig. 2 reports *B*-subjects' average payback pattern, disentangled by treatment.[25] The panel shows similar patterns in the three treatments, with no significant difference for each of the 11 hypothesized α (Kruskal-Wallis test: smallest *P-value* $= 0.346$ for $\alpha = 0.9$; Mann–Whitney test with pairwise comparisons: smallest *P-value* $= 0.165$ for $\alpha = 0.9$ in *QnoD versus QD*).

The left panel of Fig. 2 also shows that average payback patterns are increasing.[26] This follows from the prevalence of subjects whose elicited preferences display guilt aversion, as indicated by the right panel of Fig. 2.

Indeed, the right panel of Fig. 2 reports the average payback pattern of *B*-subjects disentangled by the five shapes of payback function $\xi(\alpha)$ predicted by Proposition B.1.1 in *Online Appendix B.1*: guilt prevails ($\xi(\alpha)$ increasing), guilt prevails for high

[25] When asked to fill in again the questionnaire at the end of the experiment in *QD* and *QnoD* (cf. Table 2), with very few exceptions (3/80 in *QD* and 1/40 in *QnoD*), *B*-subjects essentially confirmed the payback pattern of phase 2, in line with consistency motives (Podsakoff et al. 2003). Thus, for these two treatments in Fig. 2 we only referred to the questionnaire in phase 2. For *NoQ* we relied on the final questionnaire, the only one filled in by *B*-subjects in this treatment.

[26] For each treatment we checked for absence of framing effects on the payback pattern due to the presentation of the 11 lines of the questionnaire in reverse order in half of the experimental sessions (Mann–Whitney test, smallest *P-value* $= 0.129$ for $\alpha = 0.9$ in *QD*). This is confirmed by a similar ratio of increasing over decreasing payback patterns in each order of presentation (χ^2 test, *P-value* $= 0.276$).

α (U-shaped $\xi(\alpha)$), reciprocity prevails ($\xi(\alpha)$ decreasing), balanced motivations ($\xi(\alpha)$ constant), and selfish preferences ($\xi(\alpha) = 0$) as a separate special case of balanced motivations. With this, we find that 138/160 (86%) *B*-subjects have a payback pattern that mimics one of these five possible quasi-convex shapes of the payback function $\xi(\alpha)$; this fraction is treatment-independent (35/40 in *NoQ*, 33/40 in *QnoD*, 70/80 in *QD*).[27] Considering only the 138 *B*-subjects qualitatively captured by our model, the right panel of Fig. 2 reports, for each possible theoretical shape of $\xi(\alpha)$, their average payback pattern and the corresponding number of *B*-subjects: guilt prevails for the majority of these *B*-subjects (81/138).

Our model also allows to estimate, for each *B*-subject, the pair (G, R) that identifies *B*'s best response to each hypothesized α, i.e., his theoretical payback function $\xi(\alpha; G, R)$. The main goal of this estimation is to describe each *B*-subject by his estimated psychological type $(\hat{G}, \hat{R})$. This is a preliminary step to test the theoretical predictions of Propositions 1–4, which rely on the different regions of psychological types (G, R) in Fig. 1. We also use the estimated parameters $\hat{G}$ and $\hat{R}$ to include in one of the five categories of the right panel of Fig. 2 also the 22/160 (14%) *B*-subjects whose payback pattern does not fit any of the corresponding shapes of $\xi(\alpha)$ (for a similar method, see Costa-Gomes et al. 2001).

The best-fit response function $\hat{\xi}(\alpha) := \xi(\alpha; \hat{G}, \hat{R})$ of a given *B*-subject minimizes the sum of the squared deviations of the theoretical payback function from the payback pattern for the 11 rows of the filled-in questionnaire. Given that $\xi(\alpha; G, R)$ is not linear, we use non-linear least square estimation, with bounds given by $0 \leq G, R \leq 1000$. To account for the small size of the sample, standard deviations are given by a (non parametric) bootstrap estimation of size 10,000.[28] Across all 160 *B*-subjects, we find that 123 have $\hat{G} > 0$, 101 have $\hat{R} > 0$, and 88 have both $\hat{G} > 0$ and $\hat{R} > 0$, with no significant treatment difference in the distribution of each of the three estimated parameters (Kruskal-Wallis test, *P-value* $= 0.358$ for G, 0.760 for R).

In Table 4, we report the distribution of the 160 *B*-subjects' estimated psychological types across the five possible shapes of the corresponding payback function $\xi(\alpha)$ of Proposition B.1.1 in *Online Appendix B.1* (categories of psychological types). The number of *B*-subjects whose payback pattern is not qualitatively captured by the five predicted shapes—a total of 22/160—is reported in parentheses.[29]

Table 4 shows no significant difference between the distributions of types in *NoQ* and *QnoD* (χ^2 test, *P-value* $= 0.639$), which allows us to pool the data of these two treatments (column *NoQ-QNoD* in Table 4) so as to have the same number of observations without disclosure (*NoQ-QnoD*) and with disclosure (*QD*). Table 4 also shows no significant difference between the distributions of psychological types in

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

[28] In *Online Appendix C* we provide the non-linear least square estimates $\hat{G}$ and $\hat{R}$ (and standard deviations associated to each estimated parameter) for the 160 *B*-subjects in our experiment.

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

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

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

The table reports, for each treatment and category of psychological types: the number of *B* -subjects with elicited $(\hat{G}, \hat{R})$ in that category; within parentheses, the number of *B*-subjects with elicited $(\hat{G}, \hat{R})$ in that category, but with payback pattern not captured by the corresponding shape of $\xi(\alpha)$ in the right panel of Fig. 2. Column *NoQ-QnoD* pools the observations of *NoQ* and *QnoD*

NoQ-QNoD and *QD* (last two columns of Table 4: χ^2 test, *P-value* $= 0.734$). This is further evidence that the presence or absence of information disclosure does not affect subjects' answers to the questionnaire.

Together with the right panel of Fig. 2, Table 4 also shows that, independently of the treatment, the guilt component is prevalent for more than half of the *B*-subjects, while reciprocity prevails for only 16% of them.[30] There is also a non-negligible number of *B*-subjects (5%) for whom guilt prevails when α is high, and reciprocity prevails otherwise (U-shaped payback function). The remaining subjects have a flat estimated payback function (balanced motivations). The majority of them are selfish (0 payback regardless of α, 15% of the sample). The estimated payback function of the others (9% of the sample) is consistent with inequity aversion: these subjects aim at an interior distribution independent of α.

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

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

4.2 Behavior under disclosure of the filled-in questionnaire

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

Figure 3 reports the *observed versus predicted behavior* of matched *pairs* in phase 3 of *QD*, the only phase in our experimental design that supposedly approximates a Trust Minigame with complete information. Figure 3a refers to the three regions of the parameter space (G, R) of Fig. 1, which correspond to the complete-information predictions of rationalizability based on forward induction of Proposition 1. Figure 3b refers to the two regions of the parameter space (G, R) of Fig. 1b, which correspond to the equilibrium refinement of Proposition 2.

In both figures, for each region and for each category of psychological type from Table 4, we report in bold the number of "**classified**" *B*-subjects and in Italics the number of remaining ("*unclassified*") subjects.[31]

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

[31] Classified *B*-subjects in Fig. 3a have a $(\hat{G}, \hat{R})$ that can be assigned to one of the *three* regions of the parameter space (G, R) of Fig. 1a with a level of significance of at most 10% (P-values estimated by bootstrap). For Fig. 3b the same holds for the *two* regions of the parameter space (G, R) of Fig. 1b.

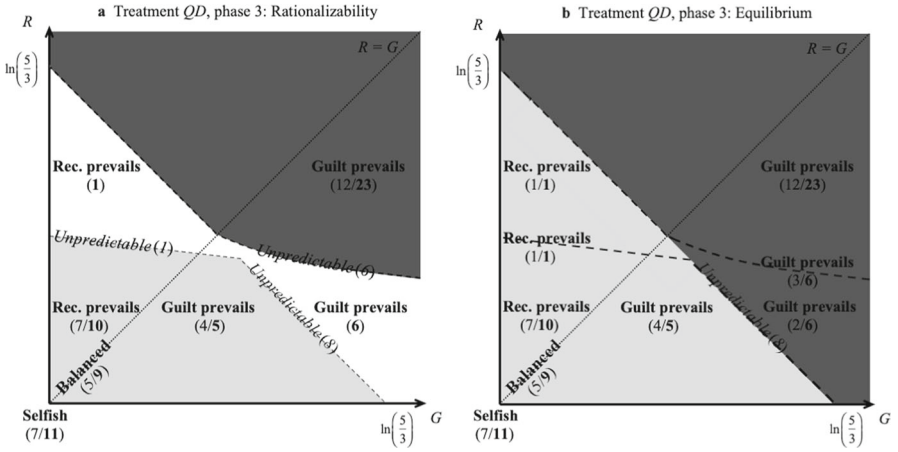

Fig. 3 Observed *versus* predicted behavior (strategy pairs) in phase 3 of *QD*. Figure 3a refers to the complete-information predictions of *rationalizability* (Proposition 1). Figure 3b refers to the complete-information *equilibrium* predictions (Proposition 2). Each ratio indicates observed (normal font) *versus* predicted (bold) behavior in phase 3 of *QD*. Number in Italics indicate unclassified *B*-subjects. Estimated types in the white (intermediate) region of Fig. 3a are classified, but do not yield a prediction according to rationalizability; thus, we do not report observed behavior

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

Predicted behavior of A–B pairs Given the estimated psychological type $(\hat{G}, \hat{R})$, we can make a prediction for about 73% (58/80) of pairs in phase 3 of *QD* according to rationalizability (bold numbers in regions $\mathbb{T}$ and $\mathbb{S}$ of Fig. 3a) and for 90% (72/80) of pairs according to the equilibrium predictions (bold numbers in Fig. 3b).

For the latter, all pairs with a *B*-subject for whom guilt does not prevail fall in the (*Dissolve*, *Take*) region. These include (as predicted) all pairs with a selfish *B*-subject, and all those with a *B*-subject who is balanced or for whom reciprocity prevails. Conversely, pairs with a *B*-subject for whom guilt prevails fall in both regions. For this reason, we can characterize the two regions according to the level of guilt sensitivity of the *B*-subject in the pair. With this, we refer to the *B*-subjects for whom guilt prevails and who are in the (*Continue*, *Share*) region of Fig. 3b as **"high-guilt" types**. Note that these are the great majority of classified pairs in the region where guilt prevails (35 *versus* 5). With this, we refer to all *B*-subjects in the (*Dissolve*, *Take*) region of Fig. 3b as **"low-guilt" types**. These low-guilt types are the 5 *B*-subjects for whom guilt prevails but is not high enough, and the above-mentioned selfish, balanced and reciprocal *B*-subjects, for whom guilt does not prevail.

Hence, we predict guilt aversion to be the main driver of the cooperative equilibrium in the Trust Minigame with complete information, as summarized in the following result:

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

Observed behavior of A–B pairs In Fig. 3a pooled ratios of observed *versus* predicted behavior in phase 3 of *QD* show a 60% (35/58) rate of success of the complete-information predictions for phase 3 of *QD*. The rate of success is not significantly different in Fig. 3b (58%, 42/72; χ^2 test, *P-value* $= 0.816$). Both rates of success are significantly higher than the one (25%) of a random guess over the four possible strategy profiles (χ^2 test, *P-value* $= 0.000$ for both Fig. 3a, b).

At first sight, our complete-information predictions seem to be more successful for pairs predicted to choose (*Dissolve*, *Take*)—66% in Fig. 3a and 68% in Fig. 3b—than for pairs predicted to choose (*Continue*, *Share*)—52% in Fig. 3a and 49% in Fig. 3b. This might be driven by significantly more pairs playing (*Dissolve*, *Take*): 47% *versus* 21% (*P-value* $= 0.003$) in Fig. 3a, 43% *versus* 25% (*P-value* $= 0.022$) in Fig. 3b. However, the two rates of success are not significantly different from each other (*P-value* $= 0.303$ in Fig. 3a, *P-value* $=0.102$ in Fig. 3b). Furthermore, if we do not consider out-of-equilibrium observed behavior, then almost 90% of (*Dissolve*, *Take*) and almost 100% of (*Continue*, *Share*) observed behavior fall in their respective prediction regions. These two fractions are not significantly different (respectively, *P-value* $= 0.159$ in Fig. 3a; *P-value* $= 0.276$ in Fig. 3b).

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

Result 3 Complete-information *rationalizability* explains 60% of the observed behavior of predicted matched pairs after questionnaire disclosure (phase 3 of treatment *QD*). Similar results are found under the complete-information *equilibrium* refinement.

In Fig. 4 we present *subjects' observed choices and beliefs* in phase 3 of *QD*, disentangled by role and by *B*'s psychological type focusing on equilibrium predictions of Fig. 3b.[32] We discuss experimental results about *A*-subjects first, and then about matched *B*-subjects.

Behavior and beliefs of A-subjects As reported in Fig. 4, *A*-subjects matched with a high-guilt *B*-subject show a significantly higher (at the 1% level) frequency of *Continue* (+50%, χ^2 test) and first-order belief α (+29% on average, Mann–Whitney test). A significant (at the 1% level) positive correlation is found between the *Continue* choice and α (rank-biserial correlation coefficient, *Somers' d* $= 0.59$).

A further result supporting the complete-information predictions is the significant (at the 1% level) positive correlation found in phase 3 of *QD* between $(\hat{G} + \hat{R})$—the feature of *B*'s estimated psychological type $(\hat{G}, \hat{R})$ relevant for the equilibrium analysis of Proposition 2—and both *A*'s choice of *Continue* ($d = 0.52$) and α (*Spearman's*

[32] We consider equilibrium rather than rationalizability predictions since, by construction, they capture a higher number of pairs in the two regions of predictions of Fig. 3. All the results below also hold if we rely on the rationalizability predictions of Fig. 3a.

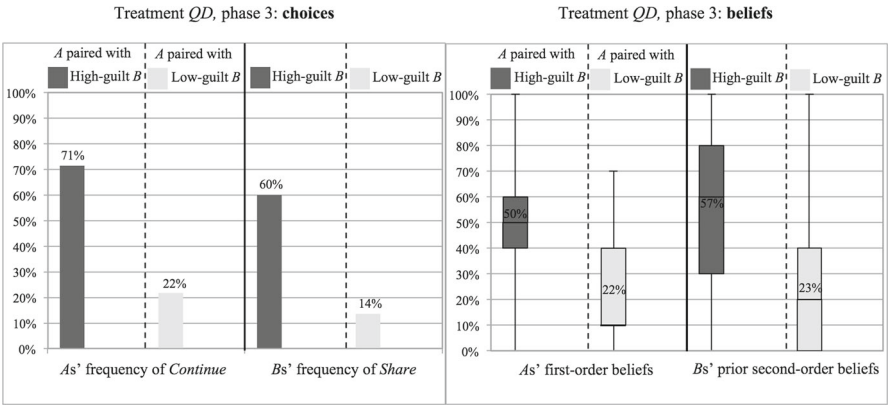

Fig. 4 *A*'s and *B*'s choices and beliefs in phase 3 of *QD*, disentangled by *B*'s type. The figure reports, for phase 3 of *QD*: on the left panel, the frequency of *A*s' *Continue* choices and of matched *B*s' *Share* choices; on the right panel, the box plot and average of *A*s' first-order belief and *B*s' unconditional second-order belief of *Share*. The color code is related to Fig. 4b: 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

$\rho = 0.44$). This is mainly due to the guilt component $\hat{G}$ ($d = 0.54$ with *Continue*, $\rho = 0.48$ with α), while for $\hat{R}$ we find a low negative correlation with both *A*-subjects' choice ($d = -0.20$, *P-value* $= 0.145$) and belief ($\rho = -0.26$, *P-value* $= 0.025$).[33]

Finally, if we disentangle the *A*-subjects in phase 3 of *QD* according to the matched (estimated) psychological type—high-guilt *versus* low-guilt—and we focus on any of the two subgroups separately, we find no significant correlation between $\hat{G} + \hat{R}$ and both the *Continue* choice and α: the largest (in absolute value) of the four correlation coefficients is 0.14 (*P-value* $= 0.550$) between $\hat{G} + \hat{R}$ and *Continue* for *A*s matched with low-guilt *B*s. This is in line with the complete-information predictions of Proposition 2 given the matched *B*'s elicited psychological types, as summarized by the following result:

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

Behavior and beliefs of B-subjects As reported in Fig. 4, high-guilt *B*-subjects show a significantly higher (at the 1% level) frequency of *Share* (+46%, χ^2 test) and unconditional second-order beliefs $\mathbb{E}_B[\widetilde{\alpha}]$ (+34% on average, Mann–Whitney test) than low-guilt ones.

As for *Share*-belief correlation, we find a strongly significant positive correlation with the first-order point-belief ($\rho = 0.44$, *P-value* $= 0.000$).[34] The correlation

[33] We verified that $\hat{G}$ and $\hat{R}$ are statistically independent ($\rho = -0.10$, *P-value* $= 0.110$). This allows us to run the correlation analysis with *A*'s choice and first-order belief for $\hat{G}$ and $\hat{R}$ separately.

[34] Recall that we only ask *B*-subjects a (coarse) feature of their first-order beliefs, i.e., whether they expect *Continue* or *Dissolve*. For ease of notation, and with an abuse of language, we refer to such reported beliefs as *B*-subjects' first-order point-belief.

between *Share* and $\mathbb{E}_B[\widetilde{\alpha}]$ is also strongly significant ($d = 0.65$, *P-value* $= 0.000$). We find the same significant correlation if we consider only *B*-subjects for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of the *conditional* second-order belief β (those with *Continue* as first-order point-belief).[35] Focusing on the latter subjects, we observe that 90% (19/21) of those classified as high-guilt types and with $\mathbb{E}_B[\widetilde{\alpha}] \geq 1/2$ choose *Share*.

Results about the positive correlation (significant at the 1% level) between $\hat{G} + \hat{R}$ and, respectively, *Share* choice ($d = 0.52$), first-order point belief ($d = 0.59$), and $\mathbb{E}_B[\widetilde{\alpha}]$ ($\rho = 0.46$) are consistent with the theoretical predictions. Moreover, since *B* is aware that his type $(\hat{G}, \hat{R})$ is disclosed to *A*, his beliefs about *A*'s behavior and beliefs move with $\hat{G} + \hat{R}$.[36]

Disentangling by type—high-guilt *versus* low-guilt—, we find no significant correlation between $\hat{G} + \hat{R}$ and *B* -subjects' choices and first- and second-order beliefs, in any of the two subgroups considered separately: the largest correlation coefficient (in absolute value) is between $\hat{G} + \hat{R}$ and the first-order point-belief of *Share* for low-guilt *B*-subjects (0.15, *P-value* $= 0.508$). This confirms the complete-information predictions: *B*'s choice depends on $G + R$ being above or below the threshold of Proposition 2, but not on its precise value.

We summarize all this in the following result.

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

4.3 Behavior without disclosure of the filled-in questionnaire

In this section, we focus on the **"incomplete-information phases,"** i.e., those phase-treatment combinations where the filled-in questionnaire is not disclosed (phase 1 of *QD*, phases 1 and 3 of *NoQ-QnoD*). In these phases, subjects play a Trust Minigame with incomplete information about *B*'s psychological type. Throughout this subsection, we provide aggregate results about the incomplete-information phases, because we do not find significant between-treatment, or within-treatment differences. In particular, due to a significant correlation in subjects' choices and beliefs across phase 1 and phase 3 of *NoQ-QnoD*, we only consider phase 3 of this treatment, which is relevant for between-treatment comparison with phase 3 of *QD* (see Sect. 4.4). Therefore, all the results in this subsection rely on pooled data of phase 1 of *QD* and of phase

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

[36] As for *A*-subjects, also for *B*-subjects we find a significant (at the 1% level) positive correlation of choices and beliefs with $\hat{G}$, and a low negative (non-significant) correlation with $\hat{R}$.

3 of *NoQ-QnoD*. We have checked that all results below hold if considering data of phase 1 rather than phase 3 of *NoQ-QnoD*.[37]

We analyze the experimental results in light of the qualitative features of the non-degenerate equilibrium described in Proposition 4 (whose statement (4) incorporates the qualitative features of the rationalizability predictions of Proposition 3):

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

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

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

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

(4) *FI-dominance* We organize *B*-subjects' choices according to the incomplete-information predictions of Proposition 3. Figure 5, built with the same method and notation as Fig. 3a, refers to the three regions of predictions in the parameter space (G, R) of Fig. 1a. Differently from Fig. 3a, due to the absence of questionnaire disclosure, Fig. 5 only refers to *B*-subjects: the left panel reports observed (normal font) *versus* predicted (bold) behavior in phase 1 of *QD*; the right panel reports the same comparison in phase 3 of *NoQ-QnoD*.

In *QD*, as in Fig. 3a, we are able to classify 65/80 *B*-subjects, and 58 out of 65 are predictable; in *NoQ-QnoD*, we classify 50/80 *B*-subjects, and 46 out of 50 are predictable. Relying on the incomplete-information predictions of Fig. 5, and considering together phase 1 of *QD* (left panel) and phase

[37] See Fig. C.2 in *Online Appendix C*, reporting observed *versus* predicted behavior in phase 1 of *NoQ-QnoD*.

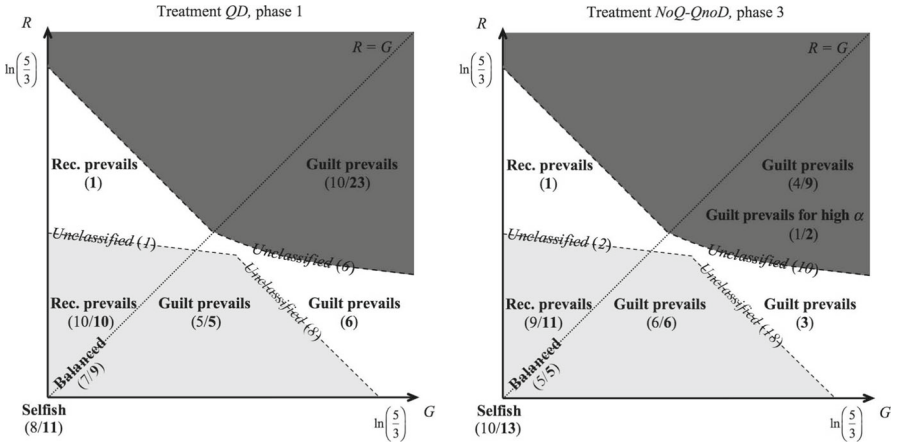

Fig. 5 Observed *versus* predicted behavior of *B*-subjects in the incomplete-information phases. The figure refers to the three regions of the parameter space (*G*,*R*) of Fig. 1a. The classification method and notation are the same as in Fig. 4a

3 of *NoQ-QnoD* (right panel), we find that *Share* is chosen by 44% of *B*-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ (dark-grey region), while it is chosen by 14% of *B*-subjects with $(\hat{G}, \hat{R}) \in \mathbb{T}$ (light-grey region), the difference being significant at the 1% level. The fact that less than half of *B*-subjects with $(\hat{G}, \hat{R}) \in \mathbb{S}$ choose *Share* seems to be mostly explained by a failure of the forward-induction inference, based on the assumption of *A* selfishness, that $\beta \geq 1/2$ (see the test of statement (2) above). Indeed, if we consider only *B*-subjects for whom $\mathbb{E}_B[\widetilde{\alpha}]$ is a rough measure of β (first-order point-belief *Continue*), we find that 88% of those with $(\hat{G}, \hat{R}) \in \mathbb{S}$ and $\mathbb{E}_B[\widetilde{\alpha}] \geq 1/2$ choose *Share*. A possible explanation is that a significant fraction of *B*-subjects believe that *A*-subjects are not selfish and may choose *Continue* even if $\alpha < 1/2$.

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

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

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

4.4 Disclosure *versus* non-disclosure of the filled-in questionnaire

We conclude with a qualitative comparison of behavior and beliefs under complete *versus* incomplete information, focusing first on A-B pairs and then on each of the two roles.

Observed behavior of A–B pairs In line with the complete-information predictions, in phase 3 of *QD* (questionnaire disclosure) there is a significant correlation ($\rho = 0.35$, *P-value* $= 0.002$) between *Continue* (resp. *Dissolve*) and *Share* (resp. *Take*); a significant correlation ($\rho = 0.33$, *P-value* $= 0.005$) is also found between the elicited values of α and $\mathbb{E}_B[\widetilde{\alpha}]$. Conversely, in non-disclosure phases (phase 1 of *QD* and phase 3 of *NoQ-QnoD*), as expected in a random-matching setting, *A*-*B* choices are independent ($\rho = -0.02$ on pooled data, *P-value* $= 0.775$) as are their beliefs ($\rho = 0.02$ on pooled data, *P-value* $= 0.819$).[38]

We then rely for all treatments on the separation criterion high-guilt *versus* low-guilt types introduced in Fig. 3b for the distribution of estimated psychological types in *QD*. As expected, the high-/low-guilt ratio for *B*-subjects in treatment *NoQ-QnoD* (24/38) is not significantly different from the *QD* treatment (35/37, *P-value* $= 0.250$; χ^2 test).[39] With this, we first compare frequencies of strategy profiles chosen by complete-information predictable *pairs* in phase 3 of *QD versus* the incomplete-information phases relevant for within-treatment and between-treatment comparisons (respectively, phase 1 of *QD* and phase 3 of *NoQ-QnoD*). Then we analyze subjects' choices and beliefs—disentangled by *role* and by *B*'s type—to make within-treatment comparisons (phase 1 *versus* phase 3 of *QD*) and between-treatment comparisons (phase 3 of *QD versus* phase 3 of *NoQ-QnoD*).[40]

In Figs. 6 and 7 we extend Fig. 4, which only refers to phase 3 of *QD*. Figure 6 shows the within-treatment comparisons of choices (frequencies) and beliefs (average and box plot) disentangled by estimated psychological type of *B* (high *versus* low-guilt). Figure 7 shows the analogous between-treatment comparisons.

Behavior and beliefs of A-subjects The controls for *A*-subjects work as they should: In each incomplete-information phase, we find no significant difference in the frequency of *Continue* and in the distribution of the first-order beliefs between *A*-subjects matched with a high-guilt type and *A*-subjects matched with a low-guilt one.[41]

Between- and within-treatment comparisons work very well for *A*-subjects matched with a *high-guilt* type: Between treatments, we find a significantly (at the 1% level) higher frequency of *Continue* (+59%, χ^2) and α (+26% on average, Mann–Whitney) in phase 3 of *QD* than in phase 3 of *NoQ-QnoD*. Within treatment, we find a similar

[38] See Table C.1 in *Online Appendix C* for an in-depth analysis.

[39] We replicated the exercise behind Fig. 3b for treatment *NoQ-QnoD* (see Fig. C.1 in *Online Appendix C*). Also there we find that all *B*-subjects in the (*Continue*, *Share*) region (24/80) are high-guilt types.

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

[41] For phase 1 of *NoQ-QnoD* this is shown in Fig. C.3 in *Online Appendix C*, where we report *A*s and *B*s' choices and beliefs in phase 1 of *QD versus NoQ-QnoD*, disentangled by *B*'s type (high-guilt *versus* low-guilt).

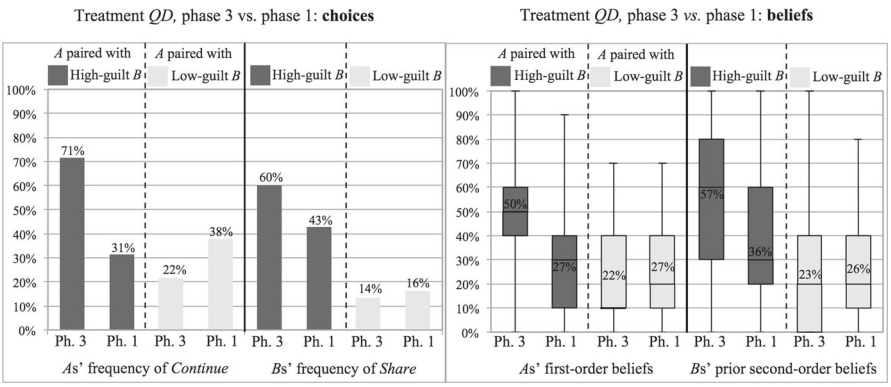

Fig. 6 *A*'s and *B*'s choices and beliefs in phase 3 *versus* phase 1 of *QD*, disentangled by *B*'s type

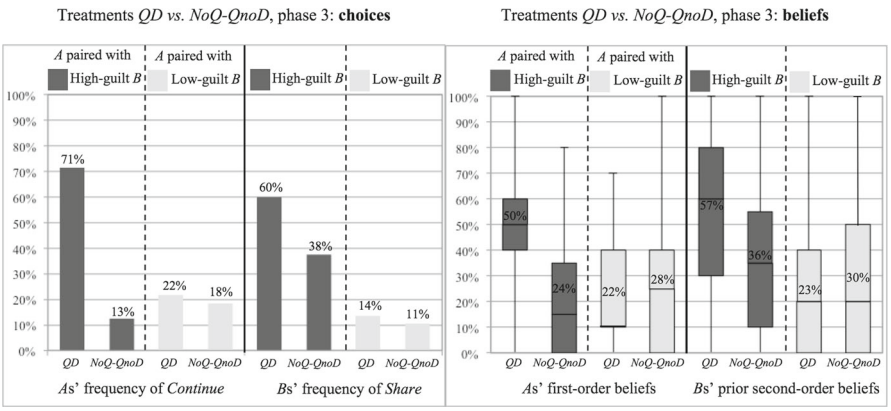

Fig. 7 *A*'s and *B*'s choices and beliefs in phase 3 of *QD versus NoQ-QnoD*, by *B*'s type

result by comparing phase 3 to phase 1 of *QD*: respectively, +40% and +23% on average, both significant at 1%.[42]

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

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

[42] A Wilcoxon matched-pairs signed-ranks test confirms the result: moving from phase 1 to phase 3 of *QD* 17/35 (*P-value* = 0.002) *A*-subjects matched with a *high-guilt* type switched from *Dissolve* to *Continue* (only 3/35 from *Continue* to *Dissolve*), and 26/35 increased (4/35 decreased, *P-value* = 0.000) their α.

As for *high-guilt B*-subjects, between- and within-treatment comparisons work quite well: Between treatments, we find a higher frequency of *Share* (+23%, χ^2 test, *P-value* $= 0.089$) and significantly higher second-order beliefs (+21% on average, Mann–Whitney test, *P-value* $= 0.012$) by comparing phase 3 of *QD* to phase 3 of *NoQ-QnoD*. Within treatment, we find similar but smaller differences by comparing phase 3 to phase 1 of *QD*: +17% (*P-value* $= 0.151$) for the frequency of *Share*, and +20% on average (*P-value* $= 0.005$) for $\mathbb{E}_B[\widetilde{\alpha}]$.[43]

Between- and within-treatment comparisons work well also for *low-guilt B*-subjects: The predicted behavior is the same under complete and incomplete information (35/37 *B*-subjects in the (*Dissolve*, *Take*) region of Fig. 3b also have $(\hat{G}, \hat{R}) \in \mathbb{T}$ in the left panel of Fig. 5), and 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. This holds regardless of whether we compare phase 3 between *QD* and *NoQ-QnoD*, or phase 3 to phase 1 within *QD* (Wilcoxon matched-pairs signed-ranks test: *P-value* $= 0.763$ for choices, *P-value* $= 0.393$ for $\mathbb{E}_B[\widetilde{\alpha}]$).

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

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

4.5 Robustness check: non-belief-dependent preferences

Our empirical characterization of subjects' psychological types is at the heart of our analysis. As a robustness check, here we consider an alternative approach based on merely distributional (i.e., non-belief-dependent) preferences. Since, in this case, truthful answers to our questionnaire should not vary across its 11 rows, any variation should be interpreted as noise. Thus, we classify subjects' answers according to the average payback (only). In what follows, we show that this alternative approach does not explain the main features of the data.

Figure 8a shows that the correlation between *B*-subjects' average payback and *Share* choices is not significant in the phase-treatment combinations with no disclosure of the payback (phase 1 of *QD* and phase 3 of *NoQ-QnoD*): $d = 0.09$, *P-value* $= 0.4212$. The correlation is significant but still low when the payback is disclosed: $d = 0.26$, *P-value* $= 0.021$. Figure 8b shows a slightly higher correlation for *A*-subjects' *Continue* choices ($d = 0.39$, *P-value* $= 0.000$), but not for their first-order beliefs ($d = 0.21$, *P-value* $= 0.065$), that under the alternative model should instead significantly increase with the average payback under disclosure. All these positive correlations are higher and more significant under our belief-dependent model, as shown in the previous data analysis (lowest $d = 0.44$, all *P-values* < 0.001). In this regard, Fig. 8 indirectly

[43] A signed-ranks test confirms the non-significant difference for choices (12/35 *versus* 6/35, *P-value* $= 0.157$) and the significant difference for unconditional second-order beliefs (25/35 *versus* 9/35, *P-value* $= 0.005$).

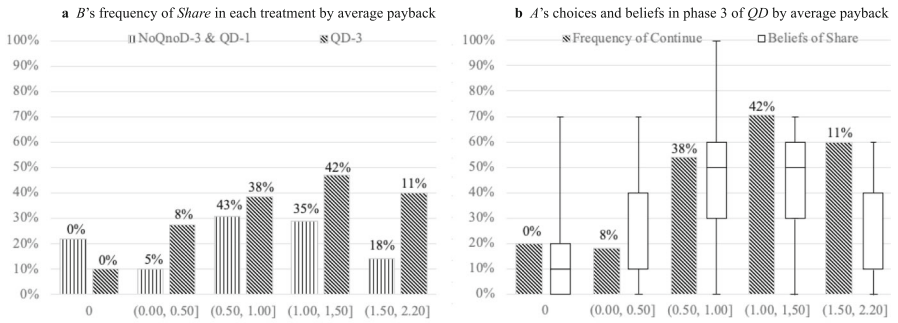

Fig. 8 *B*'s choices and *A*'s choices and beliefs, disentangled by *B*'s average payback. Figure 8a reports the frequency of *Share* choices according to *B*'s ranges of average payback disentangled by non-disclosure *versus* disclosure phase-treatment combinations. Figure 8b reports the frequency of *A*-subjects' *Continue* choices and the box plot of their first-order beliefs of *Share*. To allow direct comparison with our model, on top of each bar of the histogram we report the fraction of high-guilt types belonging to the corresponding payback range, weighted by the fraction of *B*-subjects with average payback in that range

validates our approach by showing that the highest fractions of *Continue* and *Share* are obtained for ranges of average payback to which most of the high-guilt types belong: (0.50, 1.00] without and (1.00, 1.50] with disclosure.

We can conclude that a non-belief-dependent model could only capture one feature in the data, i.e., the size of the payback (represented by $G + R$ in our model), but not how it depends on beliefs, i.e., the comparison between G and R. The prevalence of G over R explains the most important feature of our data, i.e., the positive correlation between the estimated psychological type of B and his second-order belief of *Share* under disclosure.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00199-025-01645-5.

Funding Open access funding provided by Università degli studi di Bergamo within the CRUI-CARE Agreement.

References

Attanasi, G., Nagel, R.: A survey of psychological games: theoretical findings and experimental evidence. In: Innocenti, A., Sbriglia, P. (eds.) Games, Rationality and Behavior. Essays on Behavioral Game Theory and Experiments, pp. 204–232. Palgrave McMillan, Houndmills (2008)

Attanasi, G., Battigalli, P., Manzoni, E.: Incomplete information models of guilt aversion in the trust game. Manag. Sci. **62**, 648–667 (2016)

Attanasi, G., Rimbaud, C., Villeval, M.C.: Guilt aversion in (new) games: Does partners' vulnerability matter? Games Econ. Behav. **142**, 690–717 (2023)

Attanasi, G., Battigalli, P., Manzoni, E., Nagel, R.: Belief-dependent preferences and reputation: experimental analysis of a repeated trust game. J. Econ. Behav. Organ. **167**, 341–360 (2019)

Azar, O.: The influence of psychological game theory. J. Econ. Behav. Organ. **167**, 445–453 (2019)

Bacharach, M., Guerra, G., Zizzo, D.J.: The self-fulfilling property of trust: an experimental study. Theor. Decis. **63**, 349–388 (2007)

Balafoutas, L., Fornwager, H.: The limits of guilt. J. Econ. Sci. Assoc. **3**, 137–148 (2017)

Barrett, L.F.: Solving the emotion paradox: categorization and the experience of emotion. Pers. Soc. Psychol. Rev. **10**, 20–46 (2006)

Battigalli, P., Dufwenberg, M.: Guilt in games. Am. Econ. Rev. Pap. Proc. **97**, 170–176 (2007)

Battigalli, P., Dufwenberg, M.: Dynamic psychological games. J. Econ. Theory **144**, 1–35 (2009)

Battigalli, P., Dufwenberg, M.: Belief-dependent motivations and psychological game theory. J. Econ. Lit. **60**, 833–882 (2022)

Battigalli, P., Siniscalchi, M.: Strong belief and forward induction reasoning. J. Econ. Theory **106**, 356–391 (2002)

Battigalli, P., Corrao, R., Dufwenberg, M.: Incorporating belief-dependent motivation in games. J. Econ. Behav. Organ. **167**, 185–218 (2019)

Battigalli, P., Corrao, R., Sanna, F.: Epistemic game theory without type structures: an application to psychological games. Games Econ. Behav. **120**, 28–57 (2020)

Bellemare, C., Sebald, A.: Measuring belief-dependent preferences without data on beliefs. Rev. Econ. Stud. **90**, 40–64 (2023)

Bellemare, C., Sebald, A., Strobel, M.: Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. J. Appl. Econom. **26**, 437–453 (2011)

Bellemare, C., Sebald, A., Suetens, S.: A note on testing guilt aversion. Games Econ. Behav. **102**, 233–239 (2017)

Bellemare, C., Sebald, A., Suetens, S.: Heterogeneous guilt aversion and incentive effects. Exp. Econ. **21**, 316–336 (2018)

Berg, J., Dickhaut, J., McCabe, K.: Trust, reciprocity, and social-history. Games Econ. Behav. **10**, 122–142 (1995)

Berkowitz, L., Harmon-Jones, E.: Toward an understanding of the determinants of anger. Emotion **4**, 107–130 (2004)

Bicchieri, C., Xiao, E., Muldoon, R.: Trustworthiness is a social norm, but trusting is not. Polit. Philos. Econ. **10**, 170–187 (2011)

Bracht, J., Regner, T.: Moral emotions and partnership. J. Econ. Psychol. **39**, 313–326 (2013)

Buskens, V., Raub, W.: Rational choice research on social dilemmas: embeddedness effects on trust. In: Wittek, R., Snijders, T.A.B., Nee, V. (eds.) Handbook of Rational Choice Social Research, pp. 113–150. Russell Sage, New York (2013)

Cartwright, E.: A survey of belief-based guilt aversion in trust and dictator games. J. Econ. Behav. Organ. **167**, 430–444 (2019)

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

Chao, M.: Intentions-based reciprocity to monetary and non-monetary gifts. Games **9**, 74 (2018)

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

Charness, G., Dufwenberg, M.: Participation. Am. Econ. Rev. **101**, 1213–1239 (2011)

Charness, G., Samek, A., van de Ven, J.: What is considered deception in experimental economics? Exp. Econ. **25**, 385–412 (2022)

Cooper, D.J., Kagel, J.H.: Other-regarding preferences. Handb. Exp. Econ. **2**, 217 (2016)

Costa-Gomes, M., Crawford, V.P., Broseta, B.: Cognition and behavior in normal-form games: an experimental study. Econometrica **69**, 1193–1235 (2001)

Danilov, A., Khalmetski, K., Sliwka, D.: Descriptive norms and guilt aversion. J. Econ. Behav. Organ. **191**, 293–311 (2021)
Dekel, E., Siniscalchi, M.: Epistemic game theory. In: Young, P., Zamir, S. (eds.) Handbook of Game Theory, vol. 4, pp. 619–702. North Holland, Amsterdam (2015)
Della, Lena S., Manzoni, E., Panebianco, F.: On the transmission of guilt aversion and the evolution of trust. Games Econ. Behav. **142**, 765–793 (2023)
Dhaene, G., Bouckaert, J.: Sequential reciprocity in two-player, two-stage games: an experimental analysis. Games Econ. Behav. **70**, 289–303 (2010)
Di Bartolomeo, G., Dufwenberg, M., Papa, S., Passarelli, F.: Promises, expectations & causation. Games Econ. Behav. **113**, 137–146 (2019)
Dufwenberg, M., Gneezy, U.: Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. **30**, 163–182 (2000)
Dufwenberg, M., Kirchsteiger, G.: A theory of sequential reciprocity. Games Econ. Behav. **47**, 268–298 (2004)
Dufwenberg, M., Gachter, S., Hennig-Schmidt, H.: The framing of games and the psychology of play. Games Econ. Behav. **73**, 459–478 (2011)
Dufwenberg, M., Smith, A., Van Essen, M.: Hold-up: with a vengeance. Econ. Inq. **51**, 896–908 (2013)
Ederer, F., Stremitzer, A.: Promises and expectations. Games Econ. Behav. **106**, 161–178 (2017)
Elster, J.: Emotions and economic theory. J. Econ. Lit. **36**, 47–74 (1998)
Engler, Y., Kerschbamer, R., Page, L.: Guilt averse or reciprocal? Looking at behavioral motivations in the trust game. J. Econ. Sci. Assoc. **4**, 1–14 (2018)
Falk, A., Fehr, E., Fischbacher, U.: Testing theories of fairness—intentions matter. Games Econ. Behav. **62**, 287–303 (2008)
Geanakoplos, J., Pearce, D., Stacchetti, E.: Psychological games and sequential rationality. Games Econ. Behav. **1**, 60–79 (1989)
Gómez-Miñambres, J., Schniter, E., Shields, T.W.: Investment choice architecture in trust games: when all-in is not enough. Econ. Inq. **59**, 300–314 (2021)
Guerra, G., Zizzo, D.J.: Trust responsiveness and beliefs. J. Econ. Behav. Organ. **55**, 25–30 (2004)
Harsanyi, J.: Games of incomplete information played by Bayesian players, parts I, II, III. Manag. Sci. **14**, 159–182, 320–334, 486–502 (1967–68)
Haselton, M.G., Ketelaar, T.: Irrational emotions or emotional wisdom? The evolutionary psychology of emotions and behavior. In: Forgas, J. (ed.) Hearts and Minds: Affective Influences on Social Cognition and Behavior. Frontiers of Psychology Series, pp. 21–40. Psychology Press, New York (2006)
Jensen, M.K., Kozlovskaya, M.: A representation theorem for guilt aversion. J. Econ. Behav. Organ. **125**, 148–161 (2016)
Khalmetski, K.: Testing guilt aversion with an exogenous shift in beliefs. Games Econ. Behav. **97**, 110–119 (2016)
Khalmetski, K., Ockenfels, A., Werner, P.: Surprising gifts: theory and laboratory evidence. J. Econ. Theory **159**, 163–208 (2015)
Morell, A.: The short arm of guilt? An experiment on group identity and guilt aversion. J. Econ. Behav. Organ. **166**, 332–345 (2019)
Orhun, A.Y.: Perceived motives and reciprocity. Games Econ. Behav. **109**, 436–451 (2018)
Peeters, R., Vorsatz, M.: Simple guilt and cooperation. J. Econ. Psychol. **82**, 102347 (2021)
Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., Podsakoff, N.P.: Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. **88**, 879–903 (2003)
Rabin, M.: Incorporating fairness into game theory and economics. Am. Econ. Rev. **83**, 1281–1302 (1993)
Regner, T., Harth, N.S.: Testing belief-dependent models. Working Paper, Max Planck Institute of Economics, Jena (2014)
Reuben, E., Sapienza, P., Zingales, L.: Is mistrust self-fulfilling? Econ. Lett. **104**, 89–91 (2009)
Rimbaud, C., Soldà, A.: Avoiding the cost of your conscience: belief dependent preferences and information acquisition. Exp. Econ. **27**, 491–547 (2024)
Ross, L., Greene, D., House, P.: The false consensus effect: an egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. **13**, 279–301 (1977)
Smith, A.: The Theory of Moral Sentiments. A. Millar, London (1759)
Stanca, L., Bruni, L., Corazzini, L.: Testing theories of reciprocity: Do motivations matter? J. Econ. Behav. Organ. **71**, 233–245 (2009)

Toussaert, S.: Intention-based reciprocity and signalling of intentions. J. Econ. Behav. Organ. **137**, 132–144 (2017)

Authors and Affiliations

Giuseppe Attanasi[1,5] · Pierpaolo Battigalli[2] · Elena Manzoni[3] · Rosemarie Nagel[4]

✉ Elena Manzoni
elena.manzoni@unibg.it

Giuseppe Attanasi
giuseppe.attanasi@uniroma1.it

Pierpaolo Battigalli
pierpaolo.battigalli@unibocconi.it

Rosemarie Nagel
rosemarie.nagel@upf.edu

1 Sapienza University of Rome, Rome, Italy

2 Bocconi University and IGIER, Milan, Italy

3 University of Bergamo, Bergamo, Italy

4 ICREA, Barcelona GSE, Universitat Pompeu Fabra, Barcelona, Spain

5 Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, Budapest, Hungary