Journal of Economic Behavior and Organization 167 (2019) 341–360

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game☆

Giuseppe Attanasi[a], Pierpaolo Battigalli[b], Elena Manzoni[c,*], Rosemarie Nagel[d]

[a] University of Nice Sophia Antipolis, GREDEG, France
[b] Bocconi University and IGIER, Milan, Italy
[c] University of Milan-Bicocca, Italy
[d] ICREA, Universitat Pompeu Fabra, Barcelona GSE, Spain

ARTICLE INFO

Article history:
Received 1 April 2018
Revised 8 November 2018
Accepted 3 December 2018
Available online 15 December 2018

JEL classification:
C72
C73
C91
D83

Keywords:
Repeated psychological game
Reputation
Guilt
Almost complete information

ABSTRACT

We study in a theoretical and experimental setting the interaction between belief-dependent preferences and reputation building in a finitely repeated trust game. We focus mainly on the effect of guilt aversion. In a simple two-type model, we analyze the effect of reputation building in the presence of guilt-averse trustees and derive behavioral predictions. We test these predictions in a laboratory experiment where we elicit information on trustees' belief-dependent preferences and disclose it to the paired trustor before the repeated game.

1. Introduction

Why should an agent keep an—implicit or explicit—promise or carry out a threat when this implies a material loss? Traditional game-theoretical models maintain the assumption that players are selfish and provide conditions under which repeated interaction turns such short-run losses into long-run benefits. This can work in two ways:[1] in infinitely repeated games with complete information,[2] failure to comply with an informal, or implicit agreement can trigger the play of a continuation equilibrium that is bad for the deviator; in finitely repeated games where information is incomplete, even if only slightly, deviations can trigger a costly loss of reputation.

☆ We thank Chiara Aina, Marco Boretto, Roberto Corrao, Andrea Guido, Paola Moscariello, Filippo Ricci, participants at the 9th IMEBE Conference in Madrid, the 9th SEET Workshop in Lecce, the 1st Workshop on Psychological Game Theory in Gothenburg, the 2nd Workshop on Psychological Game Theory in Norwich, and the seminar participants at University of East Anglia, University of Lyon, and Singapore Management University for useful discussions. G. Attanasi gratefully acknowledges financial support by the ERC (grant DU 283953) and the ANR 2018 GRICRIS. P. Battigalli gratefully acknowledges financial support by the ERC (grant 324219). R. Nagel gratefully acknowledges financial support by grants SEJ2005-08391, ECO2014-56154-P of the Spanish Ministry of Education, Generalitat de Catalunya, and the Barcelona Graduate School of Economics.

* Corresponding author.
E-mail address: elena.manzoni@unimib.it (E. Manzoni).

[1] See Mailath and Samuelson (2006) and the references therein.

[2] Or finitely repeated games where the stage game has multiple equilibria.

https://doi.org/10.1016/j.jebo.2018.12.002
0167-2681/

But often agents incur material losses to keep promises or carry out threats even when they are not involved in long-run relationships, because they do not only care about their material interests. A huge experimental literature suggests that preferences displaying a concern for the material payoffs of others are important to motivate agents to incur material losses so as to increase social welfare (*e.g.*, Charness and Rabin, 2002), achieve more equitable outcomes (*e.g.*, Fehr and Schmidt, 1999), or punish selfish behavior (Güth et al., 1982, and follow-up papers). Importantly, experimental evidence suggests that belief-dependent motivations play an important role. For example, the rejection of greedy offers in the Ultimatum Game is positively correlated with how much the responder initially expected to get (see references from Battigalli et al., 2015, and Aina et al., 2018), and equal sharing in Dictator and Trust Games is positively correlated with how much the chooser believes that the other agent expected to get (see Attanasi et al., 2013 and references therein; Bellemare et al., 2017; Bellemare et al., 2018, and Charness and Dufwenberg, 2006).

In this paper we study both theoretically and experimentally the interaction of reputation and belief-dependent other-regarding preferences in the context of a repeated Trust Game. Among the relevant belief-dependent preferences, we mainly focus on guilt aversion, which seems to be the prevailing motivation in the Trust Minigame (see, among others, Attanasi et al., 2013; Bellemare et al., 2017, Charness and Dufwenberg, 2006, and Cartwright 2019).[3] First, we put forward a theoretical model with role-dependent guilt aversion that highlights how incomplete information on the belief-dependent preferences of the trustee may give rise to reputation building phenomena. We adopt a multi-period model where the trustee experiences guilt (if he defects) at the end of each period of the repeated game and the trustor's disappointment refers to his expectations at the beginning of each period (cf. Battigalli et al., 2018). We show that longer cooperative paths arise the higher the trustor's prior belief on the trustee being guilt-averse due to a twofold mechanism: a higher prior belief on the trustee's guilt sensitivity in our model implies that he is indeed more likely to be a high-guilt type, and therefore more likely to choose the cooperative action. Moreover, a low-guilt type is more likely to cooperate the higher is the trustor's prior belief, due to reputation incentives. As a consequence, the trustor chooses the cooperative action more often, the higher is his prior, so that longer cooperative paths are observed. Second, we test our predictions through an experimental analysis of the four-period repeated Trust Minigame. Building on Attanasi et al. (2013), we compare a main treatment in which information on belief-dependent preferences of trustees is disclosed to their matched trustor to a control treatment in which there is no information disclosure.

The experimental literature on reputation building in repeated trust games is vast and has addressed several research questions, which are only marginally related to ours. For example, Anderhub et al. (2002) study a finitely repeated trust game with incomplete information by explicitly introducing the possibility of a trustee's type (a robot) who always feels obliged to reward trust. Engle-Warnick and Slonim (2004) compare finitely repeated trust games with partner matching with indefinitely repeated ones where in the last period of each interaction the trustor may start a new repeated game with another trustee.

There are few theoretical or experimental studies on the interaction between belief-dependent preferences and reputation building in repeated interactions. We are only aware of Balafoutas (2011), who investigates theoretically the role of guilt aversion for corruption in public administration. Corruption is modeled as the outcome of a game played between a bureaucrat, a lobby, and the public. The three-player game is assumed to be played repeatedly with an infinite horizon and a constant continuation probability.

The theoretical and experimental results in our paper contribute to the literature on belief-dependent preferences, suggesting that such preferences matter, and that they should be taken into account when designing experiments on social dilemma games. Charness and Dufwenberg (2006) and a multitude of follow-up papers suggest that either experimenters elicit and disclose information on the relevant belief-dependent preferences, or the analysis must be an incomplete-information one, a point made forcefully in Attanasi et al. (2013). Our paper shows that, if the interaction in the experimental game is repeated, disclosure of information on psychological belief-dependent preferences affects reputation building. Moreover, it suggests that belief-dependent preferences provide an alternative to (or a foundation of) the commitment types of standard reputation models.

The rest of the paper is structured as follows. Section 2 presents our theoretical model with role-dependent guilt. Section 3 describes our experimental design. Section 4 presents our behavioral predictions. Section 5 discusses our experimental results in light of the behavioral predictions. Section 6 concludes. The Appendix contains the proof of Proposition 1. An Online Appendix collects technical details about the experimental instructions.

2. The repeated Trust Minigame

In this paper, we aim to investigate the interaction between belief-dependent preferences and reputation building phenomena. We do so by focusing, as in the experimental design, on the four-period repetition of the Trust Minigame,[4] which—

[3] We model guilt aversion as the aversion to disappoint others, as in Battigalli and Dufwenberg (2007). Recent literature argues that both guilt aversion, and aversion to deviating from a descriptive norm (average behavior of others *in a similar role*) are important drivers of human behavior that may be observationally indistinguishable, *e.g.*, in social equilibria (cf. Danilov et al., 2018). The focus of our paper is, however, to test the standard psychological notion of guilt aversion within a model of reputation building. Therefore, we do not try to discriminate between guilt aversion and compliance to social norms.

[4] Our theoretical analysis can be extended to any finite repetition of the game, as can be understood from the proof of Proposition 1 (see Appendix).

Table 1
Strategic form of the Trust Minigame.

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

Table 2
Trust Minigame with role-dependent guilt.

A/*B*	*Take*	*Share*
Dissolve	1,1	1,1
Continue	$0, 4-2\theta\alpha$	2,2

in each period—is played in its simultaneous-move version. The situation of strategic interaction that constitutes the stage game of our problem is the following: Player *A* ("she") and *B* ("he") are partners on a project that has thus far yielded a total profit of € 2. Player *A* has to decide whether to *Dissolve* or to *Continue* with the partnership. If player *A* decides to *Dissolve* the partnership, the contract states that each player receives an equal share of the profit. If player *A* decides to *Continue* with the partnership, the total profit doubles (€ 4); however, in that case, player *B* has the right to decide whether to *Share* equally or *Take* the whole surplus. We call *Continue* for player *A* and *Share* for player *B* a **cooperative action**. In the simultaneous-move game of Table 1 (the strategic form of the Trust Minigame), player *B*—before knowing player *A*'s choice—has to state if he would *Take* or *Share* the higher profits.

Players' preferences over outcomes may depend on beliefs. In particular, we focus on the case in which players' preferences over outcomes depend on the co-players' beliefs (Battigalli and Dufwenberg, 2007; 2009). The belief-dependent motivations that seem to be relevant in the Trust Minigame are **guilt** and **intention-based reciprocity**. The effects of both belief-dependent motivations in the one-shot Trust Minigame have been separately analyzed by Battigalli and Dufwenberg (2009), and an experimental and theoretical analysis of the effect of preferences that display both guilt aversion and intention-based reciprocity can be found in Attanasi et al. (2013). Moving to a repeated-game setting, however, we focus on guilt aversion, as a large experimental evidence shows that guilt is indeed the dominant psychological motivation in the Trust Minigame (see, among others, Attanasi et al., 2013; Bellemare et al., 2017, and Charness and Dufwenberg 2006).

We model guilt by adapting the *simple guilt* model of Battigalli and Dufwenberg (2007) to a multi-period setting (cf. Battigalli et al., 2019). Player *B*'s guilt depends on his guilt sensitivity, θ, and on *A*'s expected disappointment, given her subjective beliefs. To analyze the effect of guilt, we need to consider the players' first- and second-order beliefs about behavior. The beliefs that will be relevant for our analysis are player *A*'s first-order belief $\alpha = \mathbb{P}_A[Share]$, and player *B*'s conditional second-order belief $\beta = \mathbb{E}_B[\tilde{\alpha}|Continue]$.[5]

Let (m_A, m_B) be the players' monetary payoffs. The **disappointment** of *A* is the difference, if positive, between *A*'s expected and actual payoffs, that is

$$D(\alpha, m_A) = \max\{0, \mathbb{E}_\alpha[\tilde{m}_A] - m_A\}.$$

A can only be disappointed after (*Continue, Take*), in which case her disappointment is $D(\alpha, m_A(Continue, Take)) = 2\alpha$. *B*'s psychological utility after (*Continue, Take*) is thus

$$u_B(m_B, m_A, \alpha, \theta) = m_B(Continue, Take) - \theta D(\alpha, m_A(Continue, Take)) = 4 - 2\theta\alpha,$$

where θ is his guilt parameter, and $\theta = 0$ means *B* is selfish (material payoff maximizer).[6] We assume that *B*'s guilt type can take two values, low or high, *i.e.*, $\theta \in \{\theta^L, \theta^H\}$, with $\theta^L = 0$ and $\theta^H > 2$. Table 2 describes the stage game with role-dependent guilt.

Note that *B* Shares if $4 - 2\theta\beta \leq 2$, that is, if $\beta \geq 1/\theta$ (with $\theta > 0$). With this, the stage game with complete information has a unique equilibrium (*Dissolve, Take*) when $\theta = \theta^L$, and two pure strategy equilibria (*Dissolve, Take*), (*Continue, Share*), when $\theta = \theta^H$.

We focus on the repeated game obtained from the four-period repetition of the psychological stage game described above. We model guilt as experienced at the end of each period, and disappointment refers to *A*'s expectations at the beginning of each period. An alternative model could consider guilt as experienced at the end of the repeated game, *i.e.*, at the end of period 4, and *A*'s disappointment being the difference between her total realized payoff and total expected payoff according to her initial belief. The two models describe different strategic interactions. To understand the difference, we refer to the work by Battigalli et al. (2018), which emphasizes the difference between periods and stages in dynamic psychological games. The main difference rests in the fact that **periods** measure the passage of time affecting players' preferences, while **stages** are merely a representation of moments in which players choose and acquire new information. A period may consist

[5] *B*'s decision depends on his belief conditional on *Continue*, because given *Dissolve* he is indifferent between *Take* and *Share*. Even if *B* does not observe *A*'s choice before moving, this conditional belief is still well defined as long as *B* assigns positive subjective probability to *Continue* (see Attanasi et al., 2016).

[6] This is a kind of state-dependent utility because *B* does not observe *A*'s belief α.

of just one stage, as in our case, or multiple stages. Our interpretation of the Trust Minigame is that each repetition of the stage game constitutes a period. As a consequence, beliefs at the beginning of period t are relevant for the computation of the expected disappointment in period t, and the intertemporal psychological utility is obtained as the sum of the one-period psychological utilities.[7] Indeed, in our experiment we elicit beliefs at the beginning of every period consistently with our interpretation of the model as a four-period repeated game rather than a one-period game with four stages. As in Battigalli et al. (2018), the difference between period and stage is relevant for the determination of A's expectation-based reference points: in a one-period game the relevant belief is the one A holds at the beginning of the first period, while in a four-period game the relevant beliefs are those A holds at the beginning of every period.

2.1. Guilt aversion and reputation building: a model

Let us now focus on the interaction between the repeated structure of the game and the incomplete information on the psychological type, which may give rise to reputation building phenomena. In the experiment we deal with a treatment in which subjects playing in role A receive information on their co-players' psychological type, and other treatments in which no information is disclosed to them. However, even when information on B's psychological type is disclosed, the situation is arguably not one of complete information but rather one of *approximately* complete information. The literature on reputation models, which started from Kreps and Wilson (1982), Kreps et al. (1982) and Milgrom and Roberts (1982), proves that in a repeated game the presence of a slight uncertainty over the opponent's type may dramatically change the set of equilibrium outcomes, for example enhancing cooperation in games where it cannot be sustained under complete information.

We build a model of repeated interaction based on the stage game of Table 2. To analyze reputation building, we assume that the guilt type of B, θ, is his private information. Player A holds a prior belief on B's type, $\mu_1 = \mathbb{P}\left[\theta = \theta^H | h_\emptyset\right]$, which is common knowledge. Varying how extreme such prior belief is allows us to compare situations in which there is almost complete information on B's type (μ_1 close to either zero or one) to situations in which there is genuine incomplete information on his psychological type (intermediate values of μ_1).

In the following proposition, which describes reputation building, we use the following notation:

- a_t: the action profile played in period t;
- $\mu_t = \mathbb{P}[\theta = \theta^H | h_{t-1}]$: A's belief about B's guilt type at the beginning of period t;
- $\sigma_t = \mathbb{P}[Share | h_{t-1}, \theta^L]$: a low-guilt B's behavior strategy at time t;
- $\gamma_t = \mathbb{P}[Continue | h_{t-1}]$: A's behavior strategy at time t.

In Proposition 1 we give an implicit description of behavior strategies showing how they depend on μ_t which in turn depends on the history h_{t-1} and on the exogenously given prior belief of player A, μ_1. The proof of the proposition is contained in the Appendix.

Proposition 1. *The following is a continuum of sequential equilibria of the four-period repetition of the psychological game in Table 2:*

A's belief on B's guilt type is

$$\mu_t = \begin{cases} \max\left\{\frac{1}{2^{5-t}}, \mu_{t-1}\right\}, & \text{if } a_{t-1} = (\cdot, Share) \text{ and } \mu_{t-1} > 0, \\ 0, & \text{if } a_{t-1} = (\cdot, Take) \text{ or } \mu_{t-1} = 0. \end{cases}$$

A high-guilt B Shares in every period.

A low-guilt B's behavior strategy is

$$\sigma_t = \begin{cases} 0, & \text{if } t = 4 \text{ or } \mu_t = 0, \\ \dfrac{\left(2^{4-t} - 1\right)\mu_t}{1 - \mu_t}, & \text{if } t < 4 \text{ and } \mu_t \in \left(0, \dfrac{1}{2^{4-t}}\right), \\ 1, & \text{if } t < 4 \text{ and } \mu_t \geq \dfrac{1}{2^{4-t}}. \end{cases}$$

Player A's behavior strategy is

$$\gamma_t = \begin{cases} 0, & \text{if } \mu_t < \frac{1}{2^{5-t}}, \\ \gamma \in (0,1), & \text{if } t = 1 \text{ and } \mu_t = \frac{1}{2^{5-t}}, \\ \frac{2}{3}\gamma_{t-1}, & \text{if } t > 1 \text{ and } \mu_t = \frac{1}{2^{5-t}}, \\ 1, & \text{if } \mu_t > \frac{1}{2^{5-t}}. \end{cases}$$

Proposition 1 describes a continuum of equilibria in which reputation building phenomena may arise, depending on A's prior belief about B's guilt aversion, μ_1. The equilibria of Proposition 1 display the traditional structure of reputation models.

First of all, a high-guilt B optimally Shares in each period, regardless of the previous history and reputation. This is in itself a relevant result, as it is obtained with a fully rational high-guilt B, and not with the assumption of a commitment

---

[7] In the language of Battigalli et al. (2015), we adopt a **slow play** model in which stages of the game correspond to periods. The alternative would be a **fast play** model in which all the stages of the game occur in the same period.

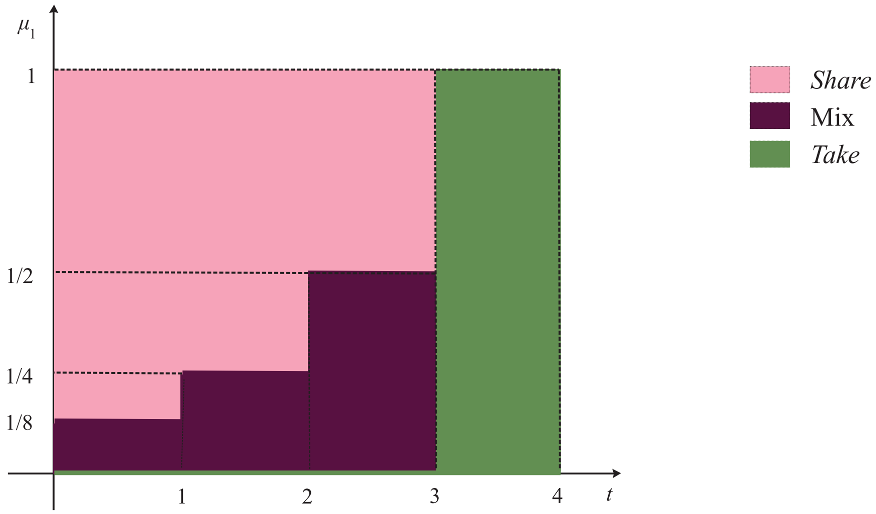

Fig. 1. Equilibrium behavior of a low-guilt B.

type. In the proof it is highlighted how in our equilibrium a high-guilt B Shares due to A choosing *Continue* only when her first-order belief in period t is $\alpha_t \geq 1/2$, which implies that B's conditional second-order belief at period t is $\beta_t \geq 1/2$.[8] A first observation that can be drawn from this analysis is therefore that belief-dependent preferences may provide an alternative to (or be the foundation of) some of the commitment types that are assumed in the standard reputation literature.

We consider equilibria with the plausible property that players do not choose weakly dominated actions; in particular, a low-guilt B always Takes in the last period, as this is weakly dominant in the stage game. In earlier periods, he either Shares, or randomizes, or Takes depending on whether his reputation in the period is respectively higher, equal, or lower than a threshold, which is increasing over time. Finally, A Continues for high values of μ_t, Dissolves for low values, and randomizes for intermediate ones.

There is however a relevant difference between our analysis and the standard reputation models, due to the type of equilibria that we consider. In the equilibrium described by Proposition 1, we have a continuum of mixed equilibria, instead of one mixed equilibrium path. This is due to our assumption that the stage game is simultaneous, together with the fact that we allow players to choose cooperative actions in $t+1$ even if A, by randomizing due to her mixed strategy, chooses *Dissolve* in period t. As a matter of fact, if B Shares in period t, μ_{t+1} is still updated correctly and is such that players can be on the mixed equilibrium path in period $t+1$. Since the stage game is such that B's action is relevant only if A Continues, also B finds it optimal to remain on the mixed equilibrium path in $t+1$ if he believes that A will do so.

We also have another kind of equilibrium, more similar to those of traditional reputation models, where as soon as A Dissolves for the first time successive play reverts to the non-cooperative profile (*Dissolve, Take*).[9] However, in the experiment, several A-subjects who defected revert to the cooperative action after observing that B-subjects Share. Thus, the equilibrium of Proposition 1 allows us to better organize the data (see Section 5.2).

Comparative statics Let us now discuss the comparative statics implied by our model.

First, we focus on the case in which B is truly a *low-guilt* player. B's reputation incentives are low for small values of μ_1,[10] and they are increasing with μ_1. As a matter of fact, the probability that B chooses the cooperative action in the first period is weakly increasing in A's prior belief, μ_1, for two reasons: if $\mu_1 > 1/8$, B Shares with probability one; moreover, for lower values of μ_1 for which the equilibrium is on the mixed path since the first period, the probability of choosing *Share* in the first period is strictly increasing in μ_1. Hence, B's incentives to build reputation through the cooperative action depend on A's belief on his type. Information disclosure on psychological types will therefore play a role in determining the equilibrium, through its effects on μ_1. Moreover, also the equilibrium path depends on A's prior belief. The higher μ_1, the higher the number of periods in which B cooperates before entering the mixed strategy path.

Fig. 1 describes the periods in which a low-guilt B Shares (pink region), Takes (green), or mixes (violet) depending on the value of μ_1. As the figure shows, the likelihood of observing longer cooperation is increasing with μ_1. Note that the

---

[8] Recall that $\beta_t = \mathbb{E}[\tilde{\alpha}_t | h_{t-1}, Continue]$ is the relevant belief for B's choice at t. Note that even a high-guilt B would choose *Take* in the last period for sufficiently low second-order beliefs. The proof of Proposition 1 in the Appendix shows how our equilibrium rests on the assumption that also the out-of-equilibrium second-order belief $\beta_t \geq 1/2$ when the probability of *Continue* is zero. This can be interpreted as a kind of forward-induction requirement.

[9] In the equilibrium where (*Dissolve, Take*) is played in every period after player A defects, A's mixing probability is pinned down precisely by the backward-induction calculation, so that we have only one equilibrium of this kind.

[10] The focus of our model is the analysis of reputation for a given (and low) number T of repetitions, so that we observe that the effects of reputation vanish when μ_1 goes to zero. The focus of Kreps and Wilson (1982) and Kreps et al. (1982) is instead on what happens if we let T become large. Their key insight is that even if μ_1 is low, there is a sufficiently large T such that reputation matters. Hence the discontinuity: with complete information behavior is independent of T, with approximately complete information behavior depends on T because of reputational concerns, as long as T is large enough.

Treatments

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

Fig. 2. Summary of the design.

fully cooperative path, defined as the path on which cooperation occurs in every period *t*, can never be observed when *B* is a low-guilt type, as he never cooperates in the last period.

Let us now consider the case in which *B* is *high-guilt* with probability μ_1. In this case, for any period $t < 4$, observing the **cooperative path up to *t***—that is, cooperation up to a certain period—is more likely the higher is μ_1. This happens in a twofold manner: first, there is a higher probability of *B* being truly high-guilt, in which case he will choose the cooperative action in every period. Second, even if *B* is low-guilt, reputation building is more profitable for him, and as a consequence he chooses the cooperative action for a larger number of periods. Hence, *A* best responds choosing *Continue* for a higher number of periods, so that longer paths of cooperation are observed for higher μ_1. Moreover, the higher μ_1, the higher the probability of a fully cooperative path.

3. The experiment

3.1. Procedures

Participants were first- and second-year undergraduate students in Economics at Bocconi University of Milan. The sessions were conducted in a computerized classroom and subjects were seated at spaced intervals. The experiment was programmed and implemented using the z-Tree software (Fischbacher, 2007).

We held 16 sessions with 20 participants per session, hence 320 subjects in total. Each person could only participate in one of these sessions. The majority of these sessions were conducted in the same time span of the experimental sessions of Attanasi et al. (2013), with none of their subjects participating in our experiment and vice versa.

Average earnings were € 15.76, including a € 5 show-up fee (minimum and maximum earnings were respectively € 6 and € 35); the average duration of a session was 65 minutes, including instructions and payment.

3.2. Design

The design is an extension of the experiment in Attanasi et al. (2013). The stage game, namely the Trust Minigame, is the one of Table 1 in Section 2. In this simultaneous-move game (the strategic form of the Trust Minigame), player *B* has to state if he would (entirely) *Take* or (equally) *Share* the higher profits *before* knowing player *A*'s choice, hence also in the case where *A* chooses *Dissolve*.

The experimental design is made of three phases and three treatments, explained in detail in Fig. 2 (for the experimental instructions see the Online Appendix). The difference among treatments depends on whether subjects playing in role *B* are asked to fill in a questionnaire in phase 2 and on whether such answers are disclosed in phase 3. We refer to these treatments, to be explained in detail below, as *No* Questionnaire (*NoQ*), Questionnaire *no* Disclosure (*QnoD*), and Questionnaire *D*isclosure (*QD*). We run 4 sessions for *NoQ* and for *QnoD* (80 subjects each) and 8 sessions for *QD* (160 subjects).

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

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. Our design coincides with the one of Attanasi et al. (2013) for phases 1 and 2, but it differs in phase 3 where it features a repeated rather than a one-shot Trust Minigame.

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

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

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* (*B*'s *unconditional second-order belief*), and to guess the choice—*Dissolve* or *Continue*—that his co-paired *A* will make (a feature of *B*'s *first-order belief*).[11]

Choice Within each pair, player *A* and player *B* simultaneously make their choice in the Trust Minigame of Table 1. In particular, player *A* selects *Dissolve* or *Continue* and player *B* selects *Take* or *Share*.

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

There are two reasons for the initial one-shot Trust Minigame of phase 1. First, a methodological reason: we want to let subjects understand the Trust Minigame and the belief-elicitation procedure before *B*-subjects, in phase 2 of *QnoD* and *QD*, fill in the questionnaire which relies on the 11 possible initial first-order beliefs of *A* (see Table 3). Second, phase 1 provides the first term of the comparison between one-shot play and repeated play needed to investigate the interplay between reputation and belief-dependent preferences.

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). *B*-subjects are asked to fill in the questionnaire (hypothetical payback scheme) of Table 3. In particular, each *B*-subject is asked to consider the following *hypothetical* situation: His new co-player *A* 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 new co-player's guess of the percentage of *B*-subjects choosing *Share*.

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

B-subjects' answers to the questionnaire of Table 3 are meant to elicit their belief-dependent preferences in treatments *QnoD* and *QD*. *B*-subjects first fill in the questionnaire on a sheet of paper and then copy the answers on the questionnaire shown on their computer screen. This is meant to make them think more carefully about their answers. *A*-subjects read and listen to the instructions of phase 2. This is made to let *A*-subjects know the task of *B*-subjects in phase 2, and in *QD* also to help them interpret the disclosed filled-in questionnaire.

In treatments *QnoD* and *QD*, it is made public information among the subjects that neither the responding *B*-subject nor anyone else will receive any payment for the answers he gives in the questionnaire of Table 3. Furthermore, in *QnoD* it is public information that *B*'s filled-in questionnaire *will not be disclosed* to anyone.

On the other hand, in *QD* it is public information that *B*'s filled-in questionnaire *will be disclosed* to a randomly-chosen *A*-subject. Actually, this subject is the one randomly matched with *B* at the beginning of phase 2. At the end of this phase, the matched *B*'s filled-in questionnaire appears on *A*'s screen, and the latter is invited to copy it on a sheet of paper, in order to increase her attention on *B*'s answers.

At this stage, subjects do not know yet that in phase 3 they are going to play again the Trust Minigame, with the same matching of phase 2. The fact that subjects are re-matched after phase 1 implies that when we say that the questionnaire is disclosed to a randomly chosen *A*-subject, we tell the truth. Yet, to separate phase 2 from phase 3 and to avoid incentives to give untruthful answers, we omit to give subjects the potentially relevant information that in phase 3 they will play the (repeated) Trust Minigame with the randomly chosen *A*-subject who received the questionnaire. While we acknowledge that we omit to give potentially relevant information, we think this is an unavoidable feature of all multistage experiments where early stages have to be isolated from later stages in order not to create an overall supergame. We emphasize that many experimental economic articles share these features.[12]

Phase 3 Phase 3—the same for all treatments—consists of a different random matching than in phase 1 (absolute-stranger matching), and of two decision tasks.

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.

The two decision tasks are an extension of those of phase 1 (and of phase 3 of the experiment in Attanasi et al., 2013). Indeed, in phase 3 subjects play the Trust Minigame in Table 1 *repeatedly* for four periods *within the same pair* (from now on, Repeated Trust Minigame). Therefore, subjects go through the two decision tasks of phase 1—belief elicitation and choice—in

[11] We elicit *B*'s unconditional second-order beliefs rather than the conditional ones, although the latter are the relevant beliefs for the correlation with the *Share* choices. We made this in order to match the 11 possible answers in *A*'s belief-elicitation task with the probability grid in the questionnaire of phase 2 (see Table 3). Thus, in order to have a manageable number of rows in the questionnaire, we only have 10 *A*-*B* pairs in each session. This is too small a number for making a reliable inference about *A*-subjects' first-order belief of *Share*, if one considers only those choosing *Continue*. Furthermore, it may occur that none out of the 10 *A*-subjects in a session choose *Continue*, which would make elicitation of *B*'s conditional second-order beliefs not incentive compatible. Note, however, that unconditional beliefs are relevant as well, as they reflect how players reason strategically before playing the Trust Minigame (see Attanasi et al., 2013).

[12] To cite only some of the articles that are most prominent or most related to our work, see, *e.g.*, the surprise re-start in the voluntary contribution game of Andreoni (1988), and the surprise disclosure of recipients' first-order beliefs in the dictator game of Ellingsen et al. (2010).

Table 3
Questionnaire (hypothetical payback scheme) in phase 2.

A's possible guesses of *Share*	Your payback (in €)
0%	Between 0.00 and 4.00
10%	Between 0.00 and 4.00
20%	Between 0.00 and 4.00
30%	Between 0.00 and 4.00
40%	Between 0.00 and 4.00
50%	Between 0.00 and 4.00
60%	Between 0.00 and 4.00
70%	Between 0.00 and 4.00
80%	Between 0.00 and 4.00
90%	Between 0.00 and 4.00
100%	Between 0.00 and 4.00

each period 1–4 of the Repeated Trust Minigame. This leads to the elicitation of 4 first-order beliefs and 4 choices for player *A*, and of 4 unconditional second-order beliefs, 4 first-order beliefs, and 4 choices for player *B*.[13]

At the end of each period, each subject within a pair is told the choice of the co-player in that period. Since in each period subjects play the simultaneous-move game of Table 1, *A*'s choice is told to *B* at the end of the period also in the case where *B*'s choice was payoff-irrelevant, *i.e., A* chose *Dissolve* in that period.

Subjects do not receive any information feedback about the belief-elicitation tasks. Indeed, at the beginning of this phase, they are informed that the gains in the belief-elicitation tasks will be communicated at the end of the experiment.

In *QnoD* and *QD*, each *B*-subject can keep with him, for the whole phase, his previously filled-in paper questionnaire. Additionally, in this phase of *QD, A* can keep the matched *B*'s filled-in questionnaire (previously copied on a sheet of paper) with her. At the beginning of phase 3 of *QD*, it is made public information that, in each pair, *B*'s filled-in questionnaire disclosed at the end of phase 2 corresponds to the matched *B*-subject of phase 3 of *QD*.[14]

Final questionnaire After phase 3, there is a final questionnaire, which is the same for all treatments (see Table 3), and equal to the one in phase 2 of *QnoD* and *QD*.

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 are allowed to give answers different from those given in phase 2.[15]

The final questionnaire provides information about *B*-subjects who did not fill in a questionnaire in phase 2 (in *NoQ*), and allows us to check whether the *B*-subjects who filled in the questionnaire in phase 2 change or confirm their answers (in *QnoD* and *QD*).[16] In the latter case, we cannot reject the hypothesis that in phase 2 subjects truthfully revealed their belief-dependent preferences.

Payment Each subject learns the co-player's choice in the Trust Minigame in phase 1, and whether her first-order belief (*A*-subject) or his first- and second-order beliefs (*B*-subject) in phase 1 and in each period of phase 3 were correct.

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

Note that, as anticipated when describing phase 1, our research question and methodology call for phase 1 and phase 3 presented in the same session on a within-subject base. With this, we are aware that our payment protocol could raise cross-phase contamination issues, but we think that this is negligible due to the asymmetry between the length of phase 1 and phase 3, with the latter paid sequentially.[17]

[13] The elicitation of beliefs at the beginning of every period is consistent with our interpretation of the model as a four-period repeated game rather than a one-period game with 4 stages. See the discussion in the first part of Section 2.

[14] Note that *A*-subjects and *B*-subjects are in the same room and read the same instructions during all the experiment. Therefore, *B*-subjects in *QD* are aware that also *A*-subjects did not know that *A*-*B* pairs with disclosed questionnaires in phase 2 would be maintained throughout phase 3. Also note that the game in phase 3 is fully described (see p. 9 of the experimental instructions in the Online Appendix), and at the beginning of the experiment subjects are told that the experiment is made of only three phases (p. 1 of the same experimental instructions). With this, we think that subjects would hardly wonder about the existence of other omitted information that may be disclosed at the end of phase 3.

[15] When *B*-subjects fill in the final questionnaire, they know that there is no further decision task to execute; therefore, they should not have any incentive to lie.

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

[17] Cox et al. (2015; Section 10.2) discuss cross-phase contamination in the context of social preferences experiments, and report such contamination when only one phase is paid randomly at the end of the experiment with no prior information (see Cox et al., 2008, and Cox, 2009). In the language of Cox et al. (2015), our payment protocol is a hybrid of a "pay all independently" (referred to phase 1 and phase 3) and a "pay all sequentially" mechanism (referred to the four periods of phase 3). Cox et al. (2015) suggest using the latter mechanism to avoid cross-phase contamination in experiments on

4. Experimental hypotheses

The model analyzed in Section 2 informs our behavioral predictions, *i.e.*, experimental hypotheses. In the setup of Proposition 1, the difference between our treatments can be interpreted as a difference in the prior belief about *B*'s type. More precisely, we expect *A*-subjects to hold high beliefs on *B* being high(low)-guilt when a filled-in questionnaire compatible with a high(low)-guilt sensitivity is disclosed—period 1 of phase 3 of treatment *QD*. On the contrary, we expect *A*-subjects to hold intermediate and more dispersed beliefs when no information is disclosed—period 1 of phase 3 of treatments *NoQ* and *QnoD*. This implies that, by comparing the treatment with approximately complete information (*i.e., QD*) to the treatments with incomplete information (*i.e., NoQ* and *QnoD*), we consider three kinds of equilibrium path: one with a particularly high μ_1, which corresponds to the case in which *A* is matched with a high-guilt *B* in phase 3 of treatment *QD*; one with intermediate values of μ_1, which corresponds to phase 3 of treatments *NoQ* and *QnoD*; and the last one with low values of μ_1, which corresponds to phase 3 of treatment *QD* for pairs in which *B* is low-guilt.

The validity of our analysis rests on one main auxiliary assumption: we assume that eliciting information does not affect subjects' behavior if the information is not disclosed.

H0.i: Subjects in treatments *NoQ* and *QnoD* show the same behavior.

A consequence of this hypothesis is that we can pool data from the two treatments that do not disclose information, *NoQ* and *QnoD*, in what we call the **'no information disclosure' (*NoQ-QnoD*) treatment**.

We introduce two other auxiliary hypotheses concerning the behavior of subjects with belief-dependent preferences different from guilt aversion. As mentioned in Section 2, there is another major type of belief-dependent preferences that may play a role in the Trust Minigame, that is, intention-based reciprocity (see Battigalli and Dufwenberg, 2009; Dufwenberg and Kirchsteiger, 2004). The theoretical analysis of Section 2 does not give predictions for reciprocal subjects. However, trustees' positive reciprocity in the Trust Game is tipically mild (see, *e.g.*, Attanasi et al., 2013), so we expect reciprocity concerns to have little effect on behavior in our experiment as well. Moreover, it can be proved that if reciprocity concerns are mild, reciprocal subjects behave as selfish ones not only in the stage game, but also in the repeated game, therefore no reputation building arises.[18] Hence, we put forward the following auxiliary hypothesis.

H0.ii Reciprocity concerned *B*-subjects behave as selfish ones.

As a natural consequence of H0.ii, we also assume that, in phase 3 of *QD, A*-subjects matched with a *B*-subject disclosing reciprocity concerns behave as those matched with a selfish *B*-subject. Hence, we put forward the following auxiliary hypothesis.

H0.iii In treatment *QD, A*-subjects matched with reciprocity concerned *B*-subjects behave as if matched with selfish ones.

If H0.ii and H0.iii are verified, the predictions we make for selfish *B*-subjects and for the corresponding matched *A-B* pairs will be applicable to reciprocity concerned ones. Notice that selfish *B*-subjects are a special case of low-guilt *B*-subjects with $\theta = 0$.

4.1. Hypotheses on A*-subjects*

We first focus on hypotheses that are only related to *A*-subjects, and, specifically, to their first-order beliefs.

As discussed above, we expect *A*-subjects to hold more polarized beliefs on their matched *B*'s type when they receive the information on the questionnaire. We also showed in Section 2.1 how, in the equilibrium of Proposition 1, a high-guilt *B* Shares more often than a low-guilt one. As a consequence, *A*-subjects matched with a high-guilt *B* in the treatment with information disclosure should hold higher first-order beliefs on their partner choosing *Share*. In the experiment, we elicit *A*'s first-order beliefs on the percentage of *B*-subjects that will choose the cooperative action. Therefore, we put forward the following hypothesis.

HA1: In *QD, A*-subjects' first-order beliefs are higher if matched with a high-guilt rather than a low-guilt or reciprocal *B*-subject.

Our theoretical model allows us to make a second prediction on the beliefs of *A*-subjects, in particular on how these beliefs vary differently over time across treatments. Few lines above, we argued how treatment *QD* is characterized by more polarized first-order beliefs of *A*-subjects on the psychological type of *B*-subjects, and how this, combined with results of Proposition 1, implies that also beliefs on the behavior of *B*-subjects are more polarized. Conversely, with incomplete information, both features of the first-order beliefs of *A*-subjects are intermediate and more dispersed. On top of this,

social preferences as those in Charness and Rabin (2002), and our Trust Minigame is strategically equivalent to some of their two-person response games. Although acknowledging the portfolio incentives of a "pay all independently" mechanism as reported by Cox et al. (2015), we emphasize that such distortion should be negligible in our payment protocol. As a matter of fact, subjects play the Trust Minigame five times, four of which (Repeated Trust Minigame of phase 3) are paid sequentially within the same pair.

[18] See Appendix B of the working paper version at ftp://ftp.igier.unibocconi.it/wp/2018/622.pdf.

Proposition 1 also characterizes the dynamic evolution of these beliefs. In the treatment with information disclosure, there is very little to learn on the opponent's type over time: if *B* is truly high-guilt, *A*'s belief on his high-guilt type is high to start with, and it remains high over time due to his cooperative behavior. If instead *B* is low-guilt, *A* assigns a very low (possibly zero) probability to *B* being high-guilt, *B* has little incentives to cooperate and build reputation, and *A*'s belief remains low. On the contrary, without information disclosure, *A*'s beliefs evolve and polarize more over time: beliefs on *B*'s guilt type evolve due to the learning that happens in equilibrium; beliefs on *B*'s behavior are affected by both learning on the guilt type and the fact that a low-guilt *B* changes behavior over time as reputation incentives fade toward the end of the game. Keeping in mind that we elicit beliefs of *A*-subjects on the behavior of *B*-subjects, we can state the following prediction.

HA2: *A*-subjects' first-order beliefs vary more over time in *NoQ-QnoD* than in *QD*. Furthermore, in *NoQ-QnoD* they are more polarized in the last than in the first period; in *QD* they are equally polarized in the last and in the first period.

4.2. Hypotheses on B-subjects

We now describe the hypotheses on beliefs and behavior of *B*-subjects. First, as discussed after Proposition 1, the equilibrium is such that low-guilt *B*-subjects display weakly higher reputation building when there is no information disclosure about their psychological type. As a matter of fact, when *B* is low-guilt, *A* holds a higher prior belief on *B*'s type in *NoQ-QnoD* than in *QD* (average prior *vs.* prior concentrated on low-guilt type). Hence, a low-guilt *B* displays more reputation building, *i.e.*, a higher frequency of *Share* in period 1 in *NoQ-QnoD* than in *QD*. This is summarized below:

HB1: Low-guilt and reciprocal *B*-subjects display more reputation building in *NoQ-QnoD* than in *QD*.

Moreover, Fig. 1 shows that, in the equilibrium of Proposition 1, reputation building, which increases the likelihood of observing *B*-subjects choosing the cooperative action, happens only when subjects are not recognized as being low-guilt (*i.e.*, when $\mu_1 > 0$). As a consequence, the following hypothesis describes the expected difference between frequencies of sharing in period 1 of the repeated game and in the one-shot Trust Minigame of phase 1. The difference is expected to be positive when reputation building matters, that is, in all cases but the one in which information on *B* being low-guilt is disclosed.

HB2: In *NoQ-QnoD*, independently of *B*'s type, for any given second-order belief of *B*, sharing is more likely in period 1 of phase 3 than in phase 1. In *QD* this is true only for high-guilt *B*-subjects.

Finally, we have a hypothesis on the link between high-guilt *B*-subjects' second-order beliefs and sharing. Our theoretical model was derived under the simplifying assumption that the guilt sensitivity of *B*-subjects is either zero or extremely high ($\theta^H > 2$). More in general, if we allow for the existence of guilt sensitivities that are high but lower than 2, we expect to observe more cooperation from subjects who hold a higher second-order belief. Moreover, as shown in the proof of Proposition 1, the threshold for sharing is lower the earlier the period, as in earlier periods a high-guilt *B* has reputation concerns on top of the underlying psychological motivations for cooperation.

HB3: In both *QD* and *NoQ-QnoD*, sharing is more likely for higher beliefs of high-guilt *B*-subjects. The threshold is lower for earlier periods.

4.3. Hypotheses on matched A-B pairs

We finally have a set of hypotheses on the frequencies of cooperative paths of matched *A*-*B* pairs. These predictions are direct implications of the comparative statics discussed after Proposition 1. The first two, HP1 and HP2, follow from the observation that longer cooperative paths are more likely the higher is *A*'s prior belief on the guilt type of *B*. The prior belief on *B* being high-guilt should be lowest when a low-guilt questionnaire is disclosed, intermediate without disclosure, and highest when a high-guilt questionnaire is disclosed. Thus, we obtain that cooperative paths up to *t* are more likely without (with) information disclosure when *B*'s type is low-guilt (high-guilt). The prediction does not hold for $t = 4$ for pairs including a low-guilt *B*, since he always Take in the last period, regardless of the treatment.

HP1: In pairs including low-guilt or reciprocal *B*-subjects, the frequencies of the cooperative path up to *t* are higher in *NoQ-QnoD* than in *QD* for $t < 4$.

HP2: In pairs including high-guilt *B*-subjects, the frequencies of the cooperative path up to *t* are higher in *QD* than in *NoQ-QnoD* for every *t*.

A last hypothesis follows from the equilibrium behavior of a pair in which *A* has a high prior belief on *B* being high-guilt, and *B* is indeed high-guilt. In this case the fully cooperative path is observed in equilibrium.

HP3: In *QD*, pairs including high-guilt *B*-subjects are on a fully cooperative path.

5. Results

We now analyze our experimental data in light of the theoretical model. In Section 5.1 we provide a categorization of *B*'s belief-dependent preferences derived from the answers to the questionnaire of Table 3. We then use this categorization in Section 5.2 to test the experimental hypotheses of Section 4.

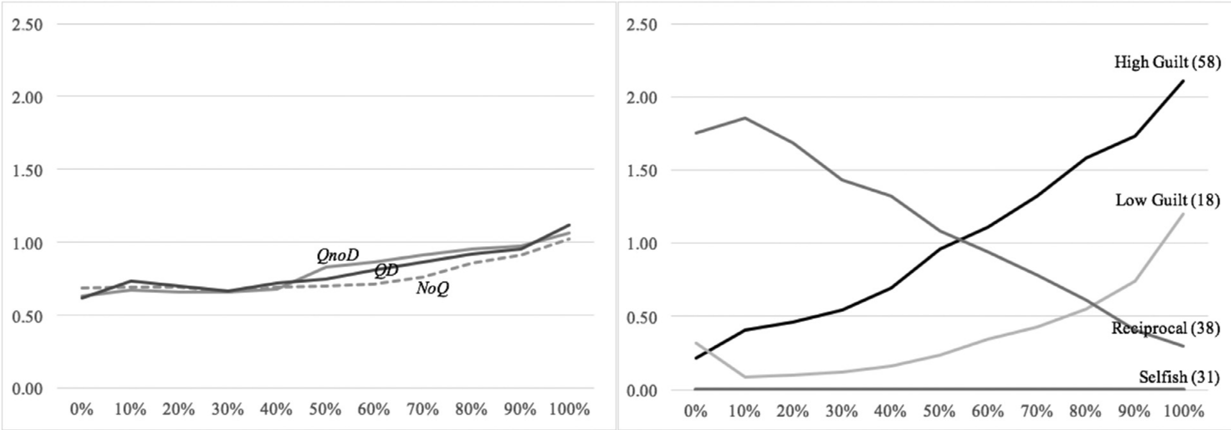

Fig. 3. *B*'s average payback pattern, by treatment (left panel) and type (right panel). The figure reports on the left panel *B*'s average payback pattern in *NoQ* (40 subjects), *QnoD* (40 subjects), and *QD* (80 subjects). On the right panel, it reports the average payback pattern of *B*-subjects according to the predicted shapes of $p(\alpha)$: high-guilt ($p(\alpha)$ increasing at least for $\alpha \geq 0.5$, and $p(1) \geq 2$), low-guilt ($p(\alpha)$ increasing at least for $\alpha \geq 0.5$, and $p(1) < 2$), reciprocal ($p(\alpha)$ decreasing at least for $\alpha \geq 0.5$), and selfish preferences ($p(\alpha) = 0$ for all α); for each average payback pattern, the intensity of the black color indicates the relative frequency of the corresponding shape in the population of *B*-subjects (reported in parentheses).

5.1. Elicitation of belief-dependent preferences through the filled-in questionnaire

As in Attanasi et al. (2013), the experimental elicitation of *B*'s belief-dependent preferences in the Trust Minigame relies on his answers to the questionnaire of Table 3. We call "**payback pattern,**" $p(\alpha)$, the actual answers of a *B*-subject, with one payback value for each hypothesized α (*A*'s belief about *B*'s action *Share*), *i.e.*, one for each of the 11 values of $\alpha \in \{0, 0.1, \ldots, 1\}$. The left panel of Fig. 3 shows *B*-subjects' average payback pattern, disentangled by treatment.

As the panel suggests, there are no treatment differences: We performed a Kruskal–Wallis test of the equality of distributions of payback values in the three treatments for each one of the 11 hypothesized α's and found the smallest *P-value* $= 0.483$ for $\alpha = 0.1$.[19] Furthermore, recall that in phase 2 of the *QnoD* and *QD* treatments *B*-subjects were asked to fill in again the questionnaire at the end of the experiment (cf. Fig. 2, final questionnaire). This was done to check whether *B*-subjects truthfully revealed their belief-dependent preferences.[20] With very few exceptions (4/40 in *QnoD*, 5/80 in *QD*), *B*-subjects confirmed the payback pattern of phase 2. Therefore, for these two treatments we only refer to the questionnaire in phase 2, while for treatment *NoQ* we rely on the final questionnaire—the only one filled in by *B*-subjects in this treatment. Finally, we checked that in each treatment there is no framing effect on the payback due to the presentation of the 11 lines of the questionnaire in reverse order in half of the experimental sessions of each treatment (Mann-Whitney test, smallest *P-value* $= 0.396$ for $\alpha = 0.2$ in *NoQ*). This is confirmed by a similar ratio of increasing over decreasing payback patterns in each order of presentation (χ^2 test, *P-value* $= 0.276$).

The left panel of Fig. 3 shows that average payback patterns are increasing. This is the result of the prevalence of subjects whose elicited preferences display guilt aversion.[21] Indeed, according to the theory of belief-dependent preferences (Battigalli and Dufwenberg, 2009), the payback pattern $p(\alpha)$ is *increasing* in α if *B* is **guilt-averse**, and *decreasing* in α if he is **reciprocal** *à la* Dufwenberg and Kirchsteiger (2004). The payback pattern of an individual with **selfish** preferences is instead constant at $p(\alpha) = 0$ for each α. Following these theoretical insights, and looking at the filled-in questionnaires, we can classify our *B*-subjects according to their elicited psychological type (we find 67/160 guilt-averse, 27/160 reciprocal, and 31/160 selfish *B*-subjects). This categorization leaves 35/160 subjects unclassified. However, we notice that in the theoretical analysis *B*'s behavior matters only when *A* Continues, and she does so only if $\alpha \geq 0.5$. As a consequence, we are able to classify some of the non-monotone payback patterns: we consider as guilt-averse also *B*-subjects with $p(\alpha)$ non-monotone in α but *increasing* for $\alpha \geq 0.5$ (9/160), and as reciprocal those with $p(\alpha)$ non-monotone in α but *decreasing* for $\alpha \geq 0.5$ (11/160). For *B*-subjects whose payback pattern is not captured by any of the above-described shapes we do not have a clear behavioral prediction, and so we categorize them as **unclassified** (15/160): the majority of them (11/15) have a positive flat payback pattern, consistent with inequity aversion.[22]

[19] A Mann-Whitney test with a pairwise comparison between treatments confirms this result (smallest *P-value* $= 0.262$ for $\alpha = 0.1$ in *QnoD vs. QD*).

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

[21] Besides Attanasi et al. (2013), which use the same elicitation method, this result is in line with other studies eliciting trustees' belief-dependent motivations in the trust game, namely Bellemare et al. (2018), and Ederer and Stremitzer (2016). Although using different elicitation methods than ours, both these studies find that the majority of trustees are guilt-averse.

[22] *B*-subjects' answers to debriefing questions about the interpretation of the filled-in questionnaire seem to confirm the categorization. These answers are available from the authors upon request.

Table 4
Categorization of *B*-subjects according to the payback pattern. The table reports, for each treatment and category of psychological types, the number of *B*-subjects with payback pattern $p(\alpha)$ in that category. Column *NoQ-QnoD* pools the observations of *NoQ* and *QnoD*.

Categories of elicited psychological types	Treatment			
	NoQ	*QnoD*	*NoQ-QnoD*	*QD*
Guilt-averse: High-guilt	14	16	30	28
Guilt-averse: Low-guilt	3	3	6	12
Reciprocal	11	9	20	18
Selfish preferences	9	8	17	14
Unclassified	3	4	7	8
TOTAL	40	40	80	80

Furthermore, we refine the categorization of guilt-averse *B*-subjects. Among all of them, we disentangle those with $p(1) \geq 2$ and those with $p(1) < 2$ and call the former **high-guilt** (58/76) and the latter **low-guilt** (18/76), so that we call high-guilt those for whom the cooperative equilibrium exists in the stage game.

All this explains the categorization in Table 4 and in the right panel of Fig. 3. The former reports the distribution of the 160 *B*-subjects' psychological types across the four possible shapes of the payback pattern $p(\alpha)$ implied by our (belief-dependent) theory-based categorization. The latter reports, for each category of psychological types, the average payback pattern and the corresponding number of classified *B*-subjects: the majority of classified *B*-subjects are guilt-averse (76/145). This fraction is treatment-independent (17/40 in *NoQ*, 19/40 in *QnoD* , 40/80 in *QD*). Table 4 shows no significant difference between the distributions of types in *NoQ* and *QnoD* (χ^2 test, *P-value* = 0.970), which allows us to pool elicited types of these two treatments (column *NoQ-QNoD* in Table 4) so as to have the same number of observations in the 'no information disclosure' (*NoQ-QnoD*) and in the information disclosure treatment (*QD*). Table 4 also shows no significant difference between the distributions of psychological types in *NoQ-QnoD* and *QD* (last two columns of Table 4: χ^2 test, *P-value* = 0.639).[23] This is further evidence that the presence or absence of information disclosure does not affect subjects' answers to the questionnaire.[24]

In the next section we will consider only classified types of Table 4 (145/160), *i.e.*, types for whom we are able to elaborate experimental hypotheses. In line with these hypotheses, we will focus on the comparison of behavior of high-guilt *vs.* low-guilt & selfish *B*-subjects. Given that selfish *B*-subjects have null guilt sensitivity, from now on we include them in the low-guilt category. We also analyze data from reciprocal *B*-subjects, pooling them together with the low-guilt ones, after checking that the relevant auxiliary hypothesis (H0.ii) is verified. We have checked that all the results below also hold if we do not consider reciprocal *B*-subjects.

5.2. Test of the experimental hypotheses

Preliminary controls Auxiliary hypothesis H0.i is verified. Considering classified types in Table 4 (37/40 for *NoQ* and 36/40 for *QnoD*), we find no significant difference between *NoQ* and *QnoD* in the behavior of *B*-subjects both in phase 1 and in each period of phase 3 (χ^2 test, smallest *P-value* = 0.407 in period 1 of phase 3). The same holds if we run the test for the three categories of classified types separately (high-guilt, low-guilt and reciprocal). This is not surprising, given that we show that there is the same distribution of classified types between *NoQ* and *QnoD* (see Table 4). We also find no significant difference in the behavior of *A*-subjects both in phase 1 and in the first three periods of phase 3 (χ^2 test, smallest *P-value* = 0.262 in phase 1).[25] Therefore, we pool the data of these two treatments (73 classified types) and we focus on the comparison between the 'no information disclosure' treatment (*NoQ-QnoD*), and *QD*, the treatment with information disclosure.

Notice that in phase 1 of the two treatments subjects are exposed to the same no-disclosure environment. Therefore, as expected, we find no between-treatment differences in the behavior of *A*-subjects and of *B*-subjects both at the aggregate

[23] Table 4 shows a lower number of low-guilt types in *NoQ-QnoD* than in *QD* (6/80 *vs.* 12/80). The difference is not significant (χ^2 test, *P-value* = 0.133), but it could raise a concern in the reader. However, if we pool together low-guilt and selfish types, which are predicted to display the same behavior, we find 23/80 in *NoQ-QnoD* and 26/80 in *QD* (χ^2 test, *P-value* = 0.607). Note that the average payback (across all values of α) of this pooled category is 0.11 in *NoQ-QnoD* and 0.18 in *QD*, with no significant treatment difference for each $\alpha \geq 0.5$ (χ^2 test, smallest *P-value* = 0.214), *i.e.*, for the subset of beliefs that are relevant for *B*'s behavior in our model and for the categorization in Table 4.

[24] Notice that our categorization is different from the one in Attanasi et al. (2013), where high-guilt *vs.* low-guilt categories are disentangled according to a non-linear least square estimation of guilt and reciprocity sensitivities, and a (non-parametric) bootstrap estimation of the probability that an elicited psychological type falls into one of the predicted regions of behavior. We have replicated Attanasi et al. (2013) categorization technique on our dataset, and found no significant difference between the distribution of elicited psychological types in Table 4 across the two categorization methods, both over the whole sample of *B*-subjects (χ^2 test, *P-value*= 0.850) and within each treatment (χ^2 test, *P-value* = 0.843 for *NoQ-QnoD* and *P-value* = 0.967 for *QD*). Estimations of guilt and reciprocity sensitivities from replication of Attanasi et al. (2013) are available from the authors upon request.

[25] The significant difference found in period 4 of phase 3 in favor of more trust by *A*-subjects in *QnoD vs. NoQ* (χ^2 test, *P*-value = 0.037) is probably due to a stronger end-game effect in treatment *NoQ*, with only 4/37 *A*-subjects choosing *Continue* (11/36 in *QnoD*) in the last period of the repeated game.

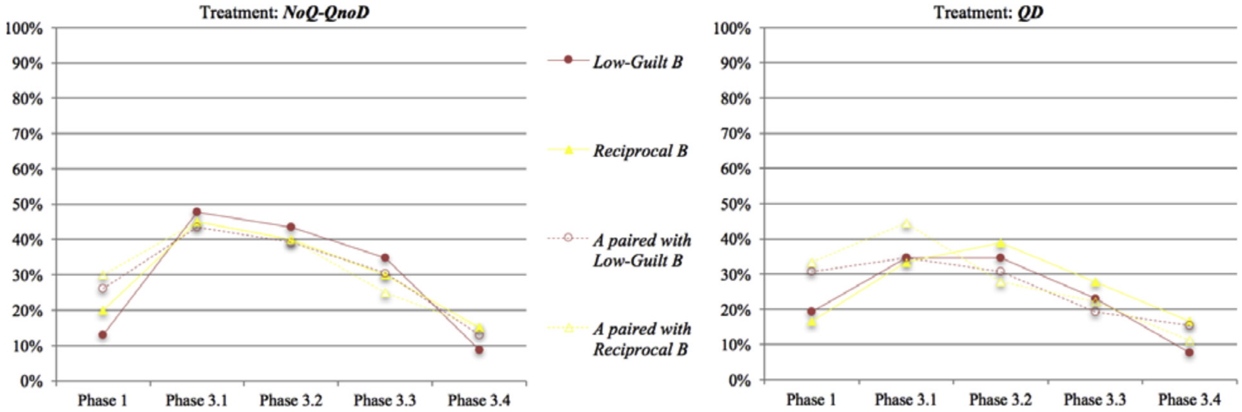

Fig. 4. Frequency of *B*'s *Share* and *A* 's *Continue* choices in low-guilt *vs.* reciprocity pairs, by treatment. Remark: To emphasize differences in frequencies within each subgroup of types, dots of phase 1 are connected to dots of phase 3.1 although game and partner change between phase 1 and phase 3.

Table 5
A-subjects' average first-order beliefs in *QD*, by matched *B*'s type, phase and period. Average beliefs in percentages; standard errors in parentheses. Results of Mann-Whitney test on equality of population medians are reported in the last row of the table.

Elicited type	Phase 1	Phase 3.1	Phase 3.2	Phase 3.3	Phase 3.4
High-Guilt	27.86	60.36	65.00	66.79	42.86
	(4.58)	(4.28)	(5.45)	(5.42)	(6.17)
Low-Guilt & Reciprocal	32.05	40.91	40.91	33.64	23.41
	(4.07)	(4.20)	(4.54)	(3.93)	(3.75)
P-value (Mann–Whitney)	0.575	0.001	0.001	0.000	0.009

level (χ^2 test, *P-value* = 0.813 for *A*s and *P-value* = 0.959 for *B*s) and if we disentangle by *B*'s type.[26] Furthermore, we find no within-treatment differences in the behavior of *A*-subjects according to the matched type—which they do not know (χ^2 test: *P-value* = 0.953 for *NoQ-QnoD* and *P-value* = 0.811 for *QD*), and a more cooperative behavior of high-guilt *B*-subjects as compared to both low-guilt *B*-subjects and reciprocal *B*-subjects, independently of the treatment (χ^2 test: in *NoQ-QnoD*, highest *P-value* = 0.062; in *QD*, highest *P-value* = 0.088).

The last control corroborates the auxiliary hypothesis H0.ii. Fig. 4 adds support and confirms that H0.ii is verified. In fact, in each treatment we find no significant difference in the behavior of low-guilt and reciprocal *B*-subjects in both phase 1 (χ^2 test: *P-value* = 0.538 for *NoQ-QnoD*, *P-value* = 0.828 for *QD*), and in each period of phase 3 (for *NoQ-QnoD*: smallest *P-value* = 0.520 in period 4; for *QD*: smallest *P-value* = 0.357 in period 4). Fig. 4 reports that the same holds for matched *A*-subjects, thus also the auxiliary hypothesis H0.iii is verified. Therefore, from now on we pool data of the two categories of psychological types (respectively in red and yellow color in Fig. 4), and perform the statistical analysis by referring to **low-guilt & reciprocal** *B*-subjects and matched pairs (in orange color in all the next figures).

All of the above also holds for subjects' beliefs.

Before we test the experimental hypotheses of Section 4, we highlight a further link between the theoretical analysis and the experimental data. The structure of the (continuum of) equilibria of Proposition 1 is such that cooperation is still possible even after *A* Dissolves on the mixed equilibrium path. This is consistent with our experimental data, showing that out of 77 occurrences of classified *A*-*B* pairs choosing (*Dissolve, Share*) in period *t* there are 24 *A*-subjects (31%) in the corresponding pairs choosing *Continue* in period $t+1$, with $t = 1, 2, 3$. This holds independently of the treatment (χ^2 test: *P-value* = 0.429).

We report below the results of the tests of the experimental hypotheses derived from our model, following the same order of presentation of Section 4, where these hypotheses are formulated.

HA1: In *QD*, *A*-subjects' first-order beliefs are higher if matched with a high-guilt rather than a low-guilt or reciprocal *B*-subject. Table 5 reports *A*-subjects' average first-order beliefs (standard errors in parentheses) according to their matched *B*'s type in treatment *QD*, for phase 1 and for each period of phase 3.

As shown by the preliminary controls, no significant difference is found for phase 1 across different matched types. As for phase 3, a significantly higher average first-order belief is recorded for *A*-subjects matched with a high-guilt type *vs.*

[26] About *A*'s choices, we find a similar fraction of *A*-subjects choosing *Continue* across *NoQ-QnoD* and *QnoD* given the type (χ^2 test: *P*-value = 0.885, 0.717 and 0.825 for *A*s matched respectively with a high-guilt *B*, a low-guilt *B*, and a reciprocal *B*). About *B*'s choices, we record a similar fraction of *B*-subjects choosing *Share* across the two treatments given the type (χ^2 test: *P*-value = 0.975, 0.559 and 0.791 respectively for a high-guilt *B*, a low-guilt *B*, and a reciprocal *B*).

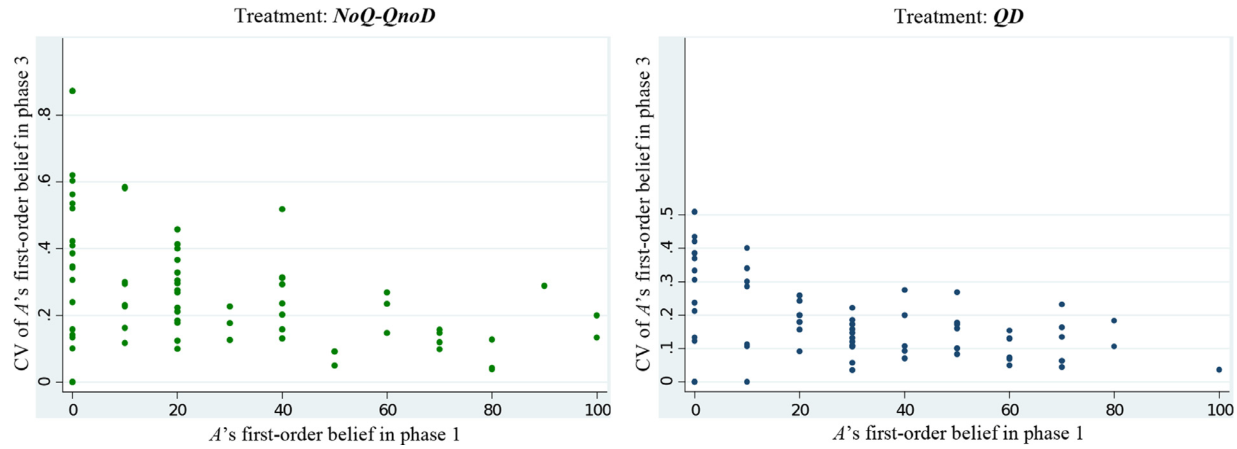

Fig. 5. Distribution of within-subject Coefficients of Variation (CV) of *A*'s first-order beliefs across periods of phase 3, by treatment and by *A*'s first-order belief (in %) in phase 1.

those matched with a low-guilt or reciprocal type. As shown in Table 5, this difference is significant at the 1% level (Mann-Whitney test) for each period of the repeated game in *QD*. Therefore, we conclude that **HA1 is verified**.

HA2: *A*-subjects' first-order beliefs vary more over time in *NoQ-QnoD* than in *QD*. Furthermore, in *NoQ-QnoD* they are more polarized in the last than in the first period; in *QD* they are equally polarized in the last and in the first period. Fig. 5 presents the distribution of *A*-subjects' coefficients of variation of first-order beliefs over the four periods of the repeated game in phase 3, disentangled by *A*-subjects' first-order belief in the one-shot game of phase 1. Each coefficient of variation is calculated on a within-subject base. The left panel refers to *NoQ-QnoD*; the right panel refers to *QD*. As Fig. 5 shows, the distribution of coefficients of variation is significantly higher in *NoQ-QnoD* than in *QD* (Mann-Whitney test, *P-value* = 0.001). This holds in particular for *A*-subjects whose first-order beliefs in phase 1 are low, *i.e.*, $\alpha < 50\%$ (*P-value* = 0.002). For these *A*-subjects, (low-guilt or reciprocal) *B*'s strategic reputation building in *NoQ-QnoD* has more room to play a positive effect on their beliefs in the first period of phase 3, thereby making them vary more over the last three periods of the repeated game. Thus, we conclude that **the first part of HA2 is verified**.

Let us now compare, for each treatment, the polarization of *A*-subjects' first-order beliefs in period 4 *vs.* period 1 of phase 3. Fig. 6 shows that beliefs are more polarized in period 4 (bottom panels) than in period 1 (top panels). However, this holds not only in *NoQ-QnoD* (left panels; Mann-Whitney: *P-value* = 0.001) but also in *QD* (right panels; *P-value* = 0.000) where instead our model predicts the same level of polarization. Thus, we conclude that **the second part of HA2 is verified in *NoQ-QnoD* but not in *QD***.

The latter result might be due to our elicitation method of *A*-subjects' first-order beliefs α_t across periods of the repeated game ($t = 1, \dots, 4$). Recall (see Section 3.2) that *A*'s elicited first-order belief is not only about the matched *B*, but about all the 10 *B*-subjects in the session; hence, we should get an elicited α_1 that is less polarized than the true one. For example, an *A*-subject who faces a low-guilt *B*-subject in period 1 of phase 3 of *QD* is asked how many of the 10 *B*-subjects in the session (the matched *B* and the other nine) will Share in that period, and she can rationally presume—despite the disclosed filled-in questionnaire of the matched low-guilt *B*—that there are some high-guilt *B*-subjects in the session. As the repeated game with paired matching unfolds, observation of no cooperation by the matched low-guilt *B* may lead *A* to decrease α_t across periods. This boosts the frequency of last-period belief $\alpha_4 = 0$ on the cooperative behavior of *B*-subjects in pairs with a disclosed low-guilt (or reciprocal) *B* in *QD*, as one might presume by looking at the bottom-right panel of Fig. 6. We come back to this point when discussing the experimental results on matched pairs.

HB1: Low-guilt and reciprocal *B*-subjects display more reputation building in *NoQ-QnoD* than in *QD*. The frequencies under scrutiny here are represented by the solid orange lines in Fig. 7. We verify that the frequency of *Share* of low-guilt & reciprocal *B*-subjects is similar in phase 1 of the two treatments (16% in *NoQ-QnoD vs.* 18% in *QD;* χ^2 test: *P-value* = 0.815). Fig. 7 shows a higher frequency of *Share* of low-guilt & reciprocal *B*-subjects in period 1 of phase 3 than in phase 1, for both treatments (47% *vs.* 16% in *NoQ-QnoD,* χ^2 test: *P-value* = 0.002; 34% *vs.* 18% in *QD, P-value* = 0.089). This increase in cooperation is due to the repeated structure of phase 3. However, in *QD*, reputation building should be toned down by disclosure of *B*-subjects' low-guilt or reciprocal type.

To analyze the negative effect of such information disclosure on reputation building, we notice that Fig. 7 also shows a higher percentage of low-guilt & reciprocal *B*-subjects choosing *Share* in the first period of phase 3 of *NoQ-QnoD* than in *QD*: 47% *vs.* 34%, although this difference is not significant (*P-value* = 0.238). The treatment difference becomes significant if we only consider those low-guilt & reciprocal *B*-subjects who switch from *Take* in phase 1 to *Share* in period 1 of phase 3, *i.e.*, those for whom reputation building is transparent: 33% in *NoQ-QnoD vs.* 16% in *QD* (χ^2 test, *P-value* = 0.070). Thus, we conclude that **HB1 is verified under some restrictions**.

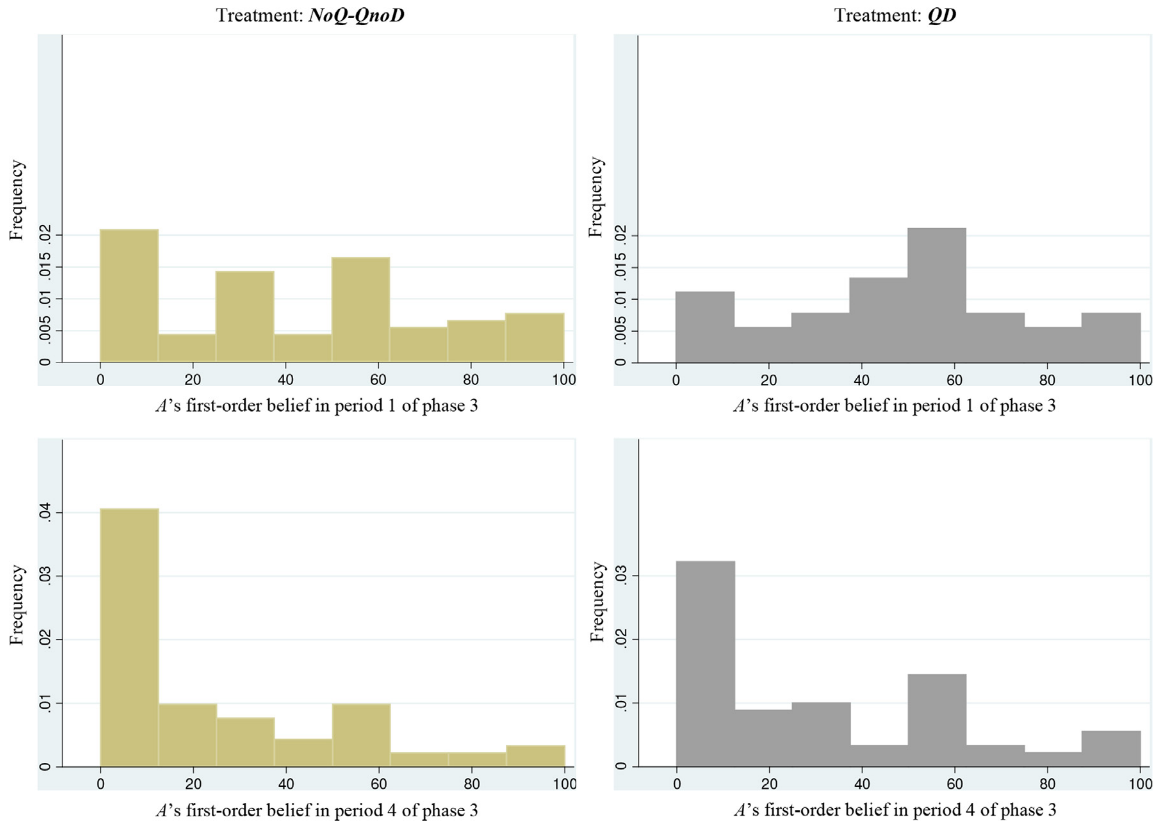

Fig. 6. Distribution of *A*'s first-order beliefs in period 1 and in period 4 of phase 3, by treatment.

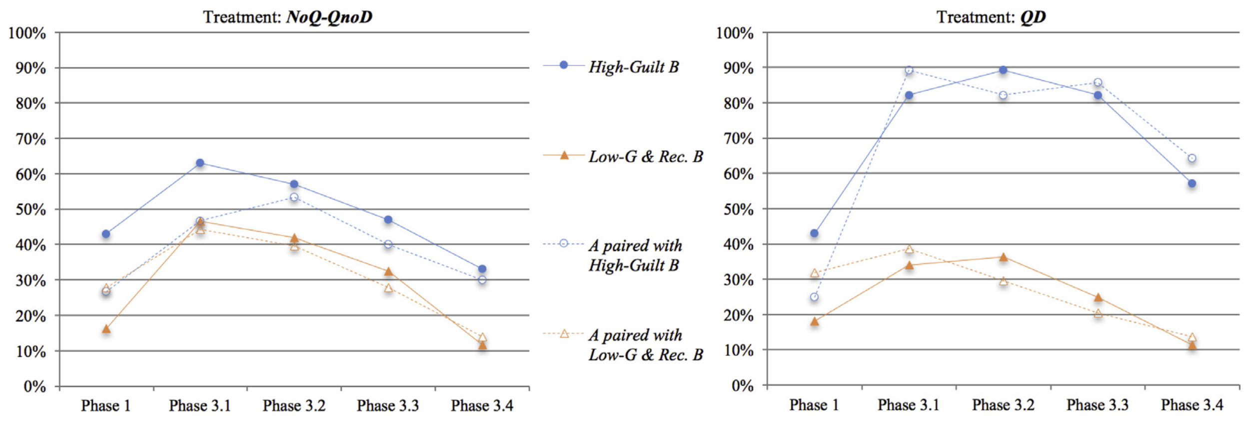

Fig. 7. Frequency of *B*'s *Share* and *A*'s *Continue* choices, by *B*'s type and by treatment. Remark: To emphasize differences in frequencies within each subgroup of types, dots of phase 1 are connected to dots of phase 3.1 although game and partner change between phase 1 and phase 3.

HB2: In *NoQ-QnoD*, independently of *B*'s type, for any given second-order belief of *B*, sharing is more likely in period 1 of phase 3 than in phase 1. In *QD* this is true only for high-guilt *B*-subjects. Fig. 7 also offers a check of this hypothesis. We first verify that the distribution of *B*-subjects' second-order beliefs in phase 1 is not significantly different across treatments (Mann-Whitney test: *P-value* = 0.699 for high-guilt; *P-value* = 0.672 for low-guilt & reciprocal), and across types within the same treatment (*P-value* = 0.896 for *NoQ-QnoD; P-value* = 0.299 for *QD*).

For what concerns the *NoQ-QnoD* treatment (left panel of Fig. 7) the hypothesis substantially finds confirmation. The difference in the frequency of *Share* choices between period 1 of phase 3 and phase 1 is significant for low-guilt & reciprocal *B*-subjects (χ^2 test, *P-value* = 0.003), and almost significant for high-guilt ones (χ^2 test, *P-value* = 0.120).

In *QD*, HB2 seems to be confirmed only for high-guilt *B*-subjects, who Share significantly more (at the 1% level) in period 1 of phase 3 than in phase 1 (χ^2 test, *P-value* = 0.003). Low-guilt & reciprocal *B*-subjects in *QD* should not be (theoretically)

Table 6
Rank correlation between *Bs*' choice and belief of *Share*, by treatment, phase and period. Correlation between a dichotomic (choice) and discrete (belief) variable is measured through Somer's *D*; * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Treatment	Phase 1	Phase 3.1	Phase 3.2	Phase 3.3	Phase 3.4
NoQ-QnoD	0.71***	0.29	0.54***	0.60***	0.84***
QD	0.68***	0.60***	0.95***	0.77***	0.86***

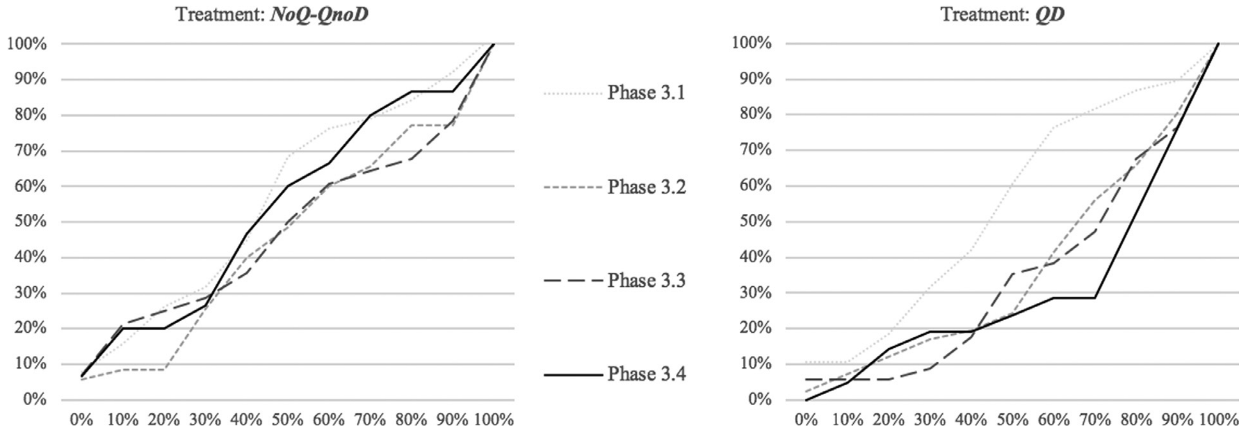

Fig. 8. Cumulative distributions of *Bs*' *Share* choices across second-order beliefs, by treatment.

more likely to Share in period 1 of phase 3 than in phase 1. Instead, the difference for low-guilt & reciprocal *B*-subjects is still significant—although only at the 10% level—(χ^2 test, *P-value* = 0.089). This may be due to the fact that some low-guilt & reciprocal *B*-subjects may actually try to build some reputation even when their questionnaire reveals information on their uncooperative type to their matched *A*-subject. Therefore, we conclude that **HB2 finds confirmation in *NoQ-QnoD* and only for high-guilt types in *QD***.

HB3: In both *QD* and *NoQ-QnoD*, sharing is more likely for higher beliefs of high-guilt *B*-subjects. The threshold is lower for earlier periods. Table 6 reports values of the rank-biserial correlation coefficient, Somers' *D*, between the *Share* choice and second-order belief of *Share*, for each treatment and each phase-period combination. A high and significant (at the 1% level) positive correlation is found in phase 1 and in each period of phase 3 for *QD*. For *NoQ-QnoD* we find the same result, except for period 1 of phase 3, possibly due to the fact that several *B*-subjects in this treatment chose *Share* in the first period of the repeated game even though they held low second-order beliefs. Indeed, the reputation-building choice of *Share* for those *B*-subjects was at no (or, at most, expectedly low) monetary cost, since—due to perfect monitoring—the matched *A*-subject would be informed about it also in the case she would choose *Dissolve* in period 1. With this, we can state that **the first part of HB3 is verified in both *QD* and *NoQ-QnoD***.

Now we test the second part of HB3. A lower threshold for second-order beliefs leading to choose *Share* in period t than for those leading to choose *Share* in period $t+1$ of the repeated game empirically gives a distribution of *Share* choices across beliefs of period $t+1$ that stochastically dominates the distribution of period t, for each $t=1,2,3$. Fig. 8 reports, for each treatment and for each period of the repeated game, the cumulative distribution of high-guilt *B*-subjects choosing *Share* across the second-order belief they hold.

For *NoQ-QnoD*, the distribution of period 1 is first-order stochastically dominated by the distribution of period 2, but the two distributions are not significantly different (Two-sample Kolmogorov–Smirnov test: *P-value* = 0.490). Furthermore, no significant difference is detected between the distributions of period 2 and period 3 (*P-value* = 0.812) and between those of period 3 and period 4 (*P-value* = 0.719). For *QD* instead it is easy to see that the distribution of period 1 is first-order stochastically dominated by the distributions of any of the other three periods (Two-sample Kolmogorov–Smirnov test: *P-value* = 0.006 for period 1 *vs.* period 2, 0.006 *vs.* period 3, 0.000 *vs.* period 4). However, both the distributions of periods 2 and 3 (*P-value* = 0.966) and the distribution of period 3 and 4 (*P-value* = 0.679) are not significantly different among them. Therefore, we conclude that **the second part of HB3 is essentially rejected**.

HP1: In pairs including low guilt or reciprocal *B*-subjects, the frequencies of the cooperative path up to *t* are higher in *NoQ-QnoD* than in *QD* for $t<4$. Fig. 9 reports, for each treatment, the frequency of (*Continue, Share*) choices of matched pairs in period t (dotted line) and the frequency of matched pairs always choosing (*Continue, Share*) up to period t (solid line), disentangled by *B*'s type in the pair. Recall that this type is disclosed in *QD* only in phase 3, and never disclosed in *NoQ-QnoD*. The controls in phase 1 work as they should: due to the random matching of *A-B* pairs, the frequency of pairs choosing (*Continue, Share*) is not significantly different between the two treatments (15% in *NoQ-QnoD vs.* 10% in *QD:* χ^2 test,

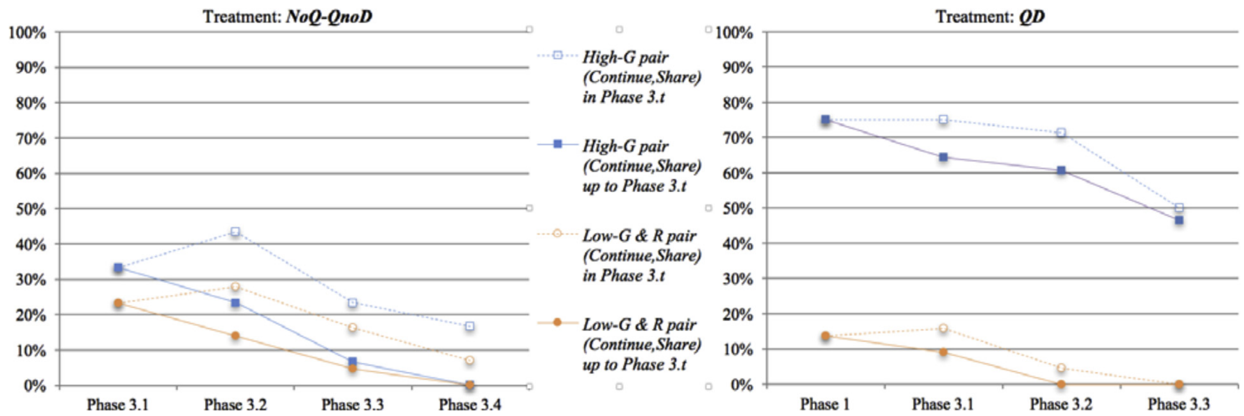

Fig. 9. Frequency of *A-B* pairs' (*Continue, Share*) choices in t and path (up to t), by *B*'s type and by treatment.

P-value = 0.329). The same holds if we disentangle pairs according to the psychological type of the *B*-subject (high-guilt *vs.* low-guilt & reciprocal).

Fig. 9 shows that for each period of phase 3 the frequency of low-guilt & reciprocal *A-B* pairs choosing (*Continue, Share*) in period t—dotted orange line—is higher in *NoQ-QnoD* than in *QD* for each $t = 1, 2, 3$. However, none of these differences is significant (χ^2 test: smallest *P-value* = 0.176 for $t = 2$). Furthermore, consistently with our theoretical predictions (see the comparative statics after Proposition 1), in period 4 of *QD* we find no low-guilt or reciprocal *A-B* pair (0/44) choosing (*Continue, Share*), with the corresponding fraction of pairs being negligible in *NoQ-QnoD* (3/43).

The same non-significant difference in favor of *NoQ-QnoD* is found if, for periods $t = 2$ and $t = 3$, we consider low-guilt & reciprocal *A-B* pairs always choosing (*Continue, Share*) up to that period—solid orange line (χ^2 test: smallest *P-value* = 0.477 for $t = 2$). In particular, neither in *NoQ-QnoD* nor in *QD* any of the pairs with a low-guilt or reciprocal *B*-subject choosing (*Continue, Share*) in period 1 was able to keep on the (*Continue, Share*) path until the end of the repeated game. This is predicted by our model (see the comparative statics after Proposition 1). Despite this, due to the absence of significant between-treatment differences, we conclude that **we find only weak support for HP1**.

HP2: In pairs including high-guilt *B*-subjects, the frequencies of the cooperative path up to *t* are higher in *QD* than in *NoQ-QnoD* for every *t*. Recall that high-guilt *B*-subjects should cooperate more than low-guilt ones in phase 1, *i.e.*, the one-shot game with no disclosure (see Attanasi et al., 2016). This control works as it should: in line with results in Attanasi et al. (2013), due to a significantly higher frequency of *Share* by high-guilt subjects, the frequency of pairs with a high-guilt *B* choosing (*Continue, Share*) is higher than the corresponding one for pairs with a low-guilt or reciprocal *B*. However, this difference is not significant in any of the two treatments (χ^2 test: *P-value* = 0.325 for *NoQ-QnoD; P-value* = 0.297 for *QD*).

Fig. 9 shows that for each period of phase 3 the frequency of pairs with a high-guilt *B* choosing (*Continue, Share*) in period t—dotted blue line—is significantly higher in *QD* than in *NoQ-QnoD* for each $t = 1, 2, 3, 4$ (χ^2 test: highest *P-value* = 0.014 for $t = 2$). The same holds if, for periods $t > 1$, we consider pairs always choosing (*Continue, Share*) up to that period—solid blue line (χ^2 test: highest *P-value* = 0.002 for $t = 2$). In particular, in *QD*, among the 14/28 pairs with a high-guilt *B* choosing (*Continue, Share*) in period 4, 13/14 had also chosen the same action profile in all previous periods of the repeated game. Conversely, in *NoQ-QnoD*, despite 10/30 pairs choosing (*Continue*, *Share*) in period 1, none of these pairs was able to keep on the (*Continue, Share*) path until the end of the repeated game. Therefore, we can conclude that **HP2 is verified**.

HP3: In *QD*, pairs including high-guilt *B*-subjects are on a fully cooperative path. Fig. 9–right panel, dotted blue lines—shows that around 70% of *A-B* pairs with a high-guilt *B* choose (*Continue, Share*) in each of the first three periods of the repeated game and 50% of them choose (*Continue, Share*) in the last period. In each period, the fraction of (*Continue, Share*) choices is significantly higher than the one (25%) obtained through a random guess over the four possible strategy profiles (χ^2 test: *P-value* < 0.001 for the first three periods, *P-value* = 0.053 for the last period). If, for each period t, we only focus on *A-B* pairs always choosing (*Continue, Share*) up to t, we find that for pairs with a high-guilt *B* this fraction is not significantly different from the previous one ((*Continue, Share*) in t) in any of the four periods (compare, respectively, solid and dotted blue lines in the right panel of Figure 9: the smallest *P-value*, 0.383, is found for period 2, χ^2 test). Therefore, almost all high-guilt pairs cooperating in a period $t > 1$ have also cooperated in all previous periods. In particular, over the 21/28 pairs with a high-guilt *B* cooperating in period 1, 13/21 (62%) cooperate until the end of the repeated game.

As a control, we check in phase 3 of *QD* that pairs with a low-guilt & reciprocal *B* are not on a (*Continue, Share*) path. Indeed, Fig. 9—right panel, dotted orange lines—shows that the fraction of these pairs choosing (*Continue, Share*) in the first two periods is not significantly different from the one of a random guess (χ^2 test: *P-value* = 0.177 for $t = 1$, *P-value* = 0.291 for $t = 2$), and in the last two periods the former fraction is significantly lower (χ^2 test: *P-value* = 0.047 for $t = 3$, *P-value* < 0.001 for $t = 4$). As highlighted above, none of these pairs chooses (*Continue, Share*) in the last period of the repeated game. This control on low-guilt & reciprocal *A-B* pairs provides further support to the fact that **HP3 is verified**.

6. Conclusions

This paper investigates the interaction between belief-dependent preferences and reputation building in a repeated Trust Minigame. More specifically, we investigate whether incomplete information on players' guilt aversion may generate reputation building phenomena in a repeated setting, both theoretically and experimentally. We focus on the four-period repetition of a Trust Minigame, and assume role-dependent guilt: *A* (the trustor) is selfish while *B* (the trustee) can feature a high or low degree of guilt aversion.

In the theoretical analysis we extend the setting of standard reputation models (Kreps et al., 1982; Kreps and Wilson, 1982; Milgrom and Roberts, 1982) to the presence of types with belief-dependent preferences. We show that reputation building phenomena may arise in presence of incomplete information on player *B*'s guilt aversion, thus enhancing cooperation.

We then analyze in a laboratory experiment the interplay between repetition of the game and information (in)completeness on *B*'s psychological type. We implement two main treatments, where subjects play the four-period Trust Minigame with partner matching. In the main treatment, subjects playing in role *A* receive information on their co-players' psychological type (information disclosure), while in the other one no information is disclosed to them. Our experimental results confirm that disclosing information on the trustee's belief-dependent preferences and thus letting players play the repeated Trust Minigame in presence of *almost complete information* affects the observed trust and cooperation behavior, leading—in the first three periods of interaction—to higher trust and cooperation than in the corresponding incomplete information when *B* is guilt-averse, and to lower trust and cooperation when he is selfish.

Our theoretical and experimental analyses emphasize once more the importance of information disclosure on belief-dependent preferences in experiments. Attanasi et al. (2013) raise the concern that experimental games where belief-dependent preferences matter are naturally played in an incomplete-information setting. They also show that in a one-shot interaction it is possible to approximate a complete-information setting by disclosing elicited information on players' belief-dependent preferences. In our paper, we take this concern for information disclosure further, and we show that in the case of repeated interactions such disclosure is not enough to approximate complete information. In the Trust Minigame, specifically, this implies that reputation building phenomena may arise, driven by the possibility that player *B* is guilt-averse. Our results highlight the need for further theoretical and experimental work on psychological games with incomplete information, and on the effects of (possibly partial) information disclosure on psychological types.

Appendix. Proof of Proposition 1

We prove Proposition 1 by backward induction. In the proof, for the sake of simplicity, we denote *A*'s actions *Continue* and *Dissolve* with respectively *C* and *D*, and *B*'s actions *Share* and *Take* with respectively *S* and *T*.

Period $t = 4$ **(last period).** A high-guilt player *B* Shares: *A* Continues only if $\alpha_4 \geq \frac{1}{2}$. Hence, if the probability that player *A* Continues is positive, then *B*'s second-order belief is $\beta_4 \geq \frac{1}{2}$. In this case, *B* finds it optimal to Share. If the probability that player *A* Continues is 0, we assume that the out-of-equilibrium conditional belief is nonetheless $\beta_4 \geq \frac{1}{2}$.[27] A low-guilt player *B* Takes: it is a (weakly) dominant action for him in the last period. Given that *B* Shares only when he is high-guilt, we have that *A*'s first-order belief is $\alpha_4 = \mu_4$. Player *A* Dissolves if $\alpha_4 < \frac{1}{2}$, Continues if $\alpha_4 > \frac{1}{2}$ and mixes if $\alpha_4 = \mu_4 = \frac{1}{2}$.

Period $t < 4$. A high-guilt player *B* Shares. To show this, first notice that in every period *t*, player *A* Continues only if $\alpha_t \geq \frac{1}{2}$, given that in the prescribed equilibrium her action does not change her information nor her future payoffs, and she mixes when $\alpha_t = \frac{1}{2}$. Therefore, if there is a positive probability that player *A* Continues, $\beta_t \geq \frac{1}{2}$. The expected utility from *Share* is:

$$\mathbb{E}\left[u_B|S,\theta^H,h_{t-1}\right] = \sum_{\tau=t}^{4}(2\gamma_\tau + (1-\gamma_\tau)) = \sum_{\tau=t}^{4}\gamma_\tau + 4 - (t-1) = \sum_{\tau=t}^{4}\gamma_\tau + 5 - t.$$

The expected utility from *Take* is:

$$\mathbb{E}\left[u_B|T,\theta^H,h_{t-1}\right] = \gamma_t\left(4 - 2\theta^H\beta_t\right) + (1-\gamma_t) + (4-t) = 3\gamma_t - 2\theta^H\beta_t\gamma_t + 5 - t.$$

Therefore, Sharing is optimal when $\sum_{\tau=t}^{4}\gamma_\tau + 5 - t \geq 3\gamma_t - 2\theta^H\beta_t\gamma_t + 5 - t$, that is when $\theta^H \geq \frac{1}{\beta_t} - \frac{\sum_{\tau=t+1}^{4}\gamma_\tau}{2\gamma_t\beta_t}$.

Given that $\frac{1}{\beta_t} - \frac{\sum_{\tau=t+1}^{4}\gamma_\tau}{2\gamma_t\beta_t} < 2$, a high-guilt *B* finds it optimal to Share. Notice that the threshold for Sharing is increasing over time: reputation concerns fade out as the end of the game approaches.

A low-guilt *B*'s expected utility if he Takes is

$$\mathbb{E}\left[u_B|T,\theta^L,h_{t-1}\right] = 4\gamma_t + (1-\gamma_t) + 4 - t = 3\gamma_t + 5 - t.$$

When *B* Takes, his reputation drops to zero in the following period ($\mu_{t+1} = 0$) and, as a consequence, the equilibrium moves to a (*D, T*) path, so that *B*'s payoff is 1 in each of the $4 - t$ remaining periods.

[27] This is an arbitrary assumption which is consistent with equilibrium analysis. We note, however, that in the last period forward induction implies $\beta_4 \geq \frac{1}{2}$ even when the probability that player *A* Continues is 0.

A low-guilt B's expected utility if he Shares is

$$\begin{aligned}\mathbb{E}\left[u_B|S,\theta^L,h_{t-1}\right] &= 2\gamma_t + (1-\gamma_t) + 4\gamma_{t+1} + (1-\gamma_{t+1}) + 4 - (t+1)\\ &= \gamma_t + 3\gamma_{t+1} + 5 - t.\end{aligned}$$

When computing the expected payoff from Sharing in period t we take into account that the play is going to be on the mixed equilibrium path, in which B is indifferent between Taking and Sharing. Hence, we can compute the expected payoff assuming that B Takes in period $t+1$. Under this assumption, the expected utility is given by the one-period expected utility from Sharing in period t (*i.e.*, $2\gamma_t + (1-\gamma_t)$), the one-period expected utility from Taking in period $t+1$ (*i.e.*, $4\gamma_{t+1} + (1-\gamma_{t+1})$), and the one-period expected utility from the (D, T) path in each of the $4-(t+1)$ remaining periods.

Hence, a low-guilt player B Shares if $\gamma_t + 3\gamma_{t+1} + 5 - t \geq 3\gamma_t + 5 - t$, that is, if $\gamma_{t+1} \geq \frac{2}{3}\gamma_t$.

If a low-guilt B Shares with probability σ_t, his reputation after Sharing is

$$\mu_{t+1} = \mathbb{P}_{\mu_t}[\theta^H|a_t = (\cdot, S)] = \frac{\mathbb{P}_{\mu_t}[a_t = (\cdot, S)|\theta^H]\mathbb{P}_{\mu_t}[\theta^H]}{\mathbb{P}_{\mu_t}[a_t = (\cdot, S)]} = \frac{\mu_t}{\mu_t + (1-\mu_t)\sigma_t}.$$

Recall that, as argued above, in every period t, player A Continues only if $\alpha_t \geq \frac{1}{2}$, mixes when $\alpha_t = \frac{1}{2}$, and Dissolves otherwise. A's first-order belief depends on her belief on B's type (μ_t) and on B's strategy (σ_t) with $\alpha_t = \mu_t + (1-\mu_t)\sigma_t$, so that in every period there is a value of μ_t for which A mixes between C and D. In period t, the mixing probability of a low-type B, σ_t, is such to induce a μ_{t+1} that implies $\alpha_{t+1} = \frac{1}{2}$. To ease notation, let $r_{t+1} = \frac{1}{\mu_{t+1}}$ denote the inverse of the reputation value μ_{t+1} that yields $\alpha_{t+1} = \frac{1}{2}$. With this, B mixes with probability σ_t such that

$$\mu_{t+1} = \frac{\mu_t}{\mu_t + (1-\mu_t)\sigma_t} = \frac{1}{r_{t+1}},$$

that is $\sigma_t = \frac{\mu_t(r_{t+1}-1)}{1-\mu_t}$.

Player A's first-order belief α_t becomes

$$\alpha_t = \mu_t + (1-\mu_t)\sigma_t = \mu_t + (1-\mu_t)\frac{\mu_t(r_{t+1}-1)}{(1-\mu_t)} = r_{t+1}\mu_t.$$

Hence, $\alpha_t = \frac{1}{2}$ when $\mu_t = \frac{1}{2r_{t+1}}$, that is, $r_t = 2r_{t+1}$. Given that $r_4 = 2$, we have that $\alpha_t = \frac{1}{2}$ when $\mu_t = \frac{1}{2^{5-t}}$, and that $\sigma_t = \frac{(2^{4-t}-1)\mu_t}{1-\mu_t}$. ■

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2018.12.002.

References

Aina, C., Battigalli, P., Gamba, A., 2018. Frustration and anger in the ultimatum game: an experiment. IGIER Working Paper no. 621.
Anderhub, V., Engelmann, D., Guth, W., 2002. An experimental study of the repeated trust game with incomplete information. J. Econ. Behav. Org. 48, 197–216.
Andreoni, J., 1988. Why free ride?: Strategies and learning in public goods experiments. J. Public Econ. 37, 291–304.
Attanasi, G., Battigalli, P., Manzoni, E., 2016. Incomplete information models of guilt aversion in the trust game. Manage. Sci. 62, 648–667.
Attanasi, G., Battigalli, P., Nagel, R., 2013. Disclosure of belief-dependent preferences in a trust game. IGIER Working Paper no. 506.
Balafoutas, L., 2011. Public beliefs and corruption in a repeated psychological game. J. Econ. Behav. Org. 78, 51–59.
Battigalli, P., Corrao, R., Dufwenberg, M., 2018. Incorporating belief-dependent motivation in games. Typescript, Bocconi University.
Battigalli, P., Corrao, R., Dufwenberg, M., 2019. Incorporating Belief-Dependent Motivation in Games. J. Econ. Behav. Org. doi:10.1016/j.jebo.2019.04.009.
Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Review (P&P) 97, 170–176.
Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.
Battigalli, P., Dufwenberg, M., Smith, A., 2015. Frustration and anger in games. IGIER Working Paper no. 539.
Bellemare, C., Sebald, A., Suetens, S., 2017. A note on testing guilt aversion. Games Econ. Behav. 102, 233–239.
Bellemare, C., Sebald, A., Suetens, S., 2018. Heterogeneous guilt aversion and incentive effects. Experimental Economics 21, 316–336 https://link.springer.com/article/10.1007/s10683-017-9543-2.
Cartwright, E., 2019. A survey of belief-based guilt aversion in trust and dictator games. J. Econ. Behav. Org. doi:10.1016/j.jebo.2018.04.01.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579–1601.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117, 817–869.
Cox, J.C., 2009. Trust and reciprocity: implications of game triads and social contexts. New Zealand Econ. Pap. 43, 89–104. Special Issue: Laboratory Experiments in Economics, Finance and Political Science.
Cox, J.C., Sadiraj, K., Sadiraj, V., 2008. Implications of trust, fear, and reciprocity for modeling economic behavior. Exp. Econ. 11, 1–24.
Cox, J.C., Sadiraj, V., Schmidt, U., 2015. Paradoxes and mechanisms for choice under risk. Exp. Econ. 18, 215–250.
Danilov, A., Khalmetski, K., Sliwka, D., 2018. Norms and guilt, CESifo working paper series 6999. CESifo Group Munich.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav 47, 268–298.
Ederer, F., Stremitzer, A., 2016. Promises and expectations. Cowles Foundation Discussion Paper No. 1931.
Ellingsen, T., Johannesson, M., Tjøtta, S., Torsvik, G., 2010. Testing guilt aversion. Games Econ. Behav. 68, 95–107.
Engle-Warnick, J., Slonim, R.L., 2004. The evolution of strategies in a repeated trust game. J. Econ. Behav. Org. 55, 553–573.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. Q. J. Econ. 114, 817–868.
Fischbacher, U., 2007. Z-tree: Zurich toolbox for readymade economic experiments. Exp. Econ. 10, 171–178.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Org. 3, 367–388.

Kreps, D.M., Milgrom, P.R., Roberts, J., Wilson, R.J., 1982. Rational cooperation in the finitely repeated prisoners' dilemma. J. Econ. Theory 27, 245–252.
Kreps, D.M., Wilson, R.J., 1982. Reputation and imperfect information. J. Econ. Theory 27, 253–279.
Mailath, G.J., Samuelson, L., 2006. Repeated games and reputations: long-run relationships. Oxford University Press, Oxford, UK.
Milgrom, P.R., Roberts, J., 1982. Predation, reputation, and entry deterrence. J. Econ. Theory, 27, 280–312.
Podsakoff, P.M., Kenzie, S.B.M., Lee, J.Y., Podsakoff, N.P., 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. 88, 879–903.