Guilt Aversion in (New) Games: Does Partners' Vulnerability Matter?

Giuseppe Attanasi, Claire Rimbaud, Marie Villeval

Guilt Aversion in (New) Games: Does Partners' Vulnerability Matter?

Giuseppe Attanasi[a], Claire Rimbaud[b], Marie Claire Villeval[c]

March 25, 2022

Abstract: We investigate whether a player's guilt aversion is modulated by the co-players' vulnerability or whether it is only activated by the willingness to avoid disappointing them. We also explore whether the nature of vulnerability (ex-post vs. ex-ante) matters. Ex-post vulnerability arises when a player's material payoff depends on another player's action (e.g., recipients in a dictator games). Ex-ante vulnerability arises when her initial endowment can be entrusted to another player (e.g., trustors in trust games). Treatments vary whether trustees can condition their decision on the belief of another player who is ex-post and/or ex-ante vulnerable. We find that trustees' guilt aversion is insensitive to the nature of the co-player's vulnerability and to the role of the co-player. Guilt is activated even absent vulnerability of co-players. It is mainly triggered by the willingness to respond to others' expectations, regardless of their responsibility or the kindness of their intentions.

JEL codes: C72, C91, D91

Keywords: Guilt Aversion, Vulnerability, Psychological Game Theory, Dictator Game, Trust Game, Experiment

[a] Sapienza Università di Roma, Dipartimento di Economia e Diritto, Via del Castro Laurenziano, 9 00161 Roma. E-mail: giuseppe.attanasi@uniroma1.it
[b] University of Innsbruck, Department of Public Finance, Universitätsstrasse 15/4, 6020 Innsbruck, Austria. E-mail: claire.rimbaud@uibk.ac.at
[c] Univ Lyon, CNRS, GATE UMR 5824, 93 Chemin des Mouilles, F-69130, Ecully, France. IZA, Bonn, Germany. E-mail: villeval@gate.cnrs.fr

Acknowledgements: We are grateful to G. Andrighetto, M. Dufwenberg, A. Guido, E. Manzoni, S. Papa, and L. Tummolini for very valuable feedback. We thank also participants at the 3rd Workshop on Psychological Game Theory (Soleto), the Workshop on Ethics and Emotions (Paris), the 1st CoCoLab Workshop (Nice), the ASFEE Conference (Toulouse), the ESA World Conference (Vancouver), and at seminars at the University of Innsbruck and the Laboratory of Agent-Based Social Simulation ISTC-CNR (Rome) for very useful comments. This project has received funding from IDEXLYON at Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and from the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French Agence Nationale de la Recherche (ANR). It has also benefited from the ANR grant GRICRIS (ANR-18-CE26-0018-01).

1 Introduction

Based on psychological insights (Baumeister et al., 1994), economists have modelled how guilt can influence actions. Within the framework of psychological game theory[1], Battigalli and Dufwenberg (2007) define guilt aversion as a belief-dependent motivation: an agent suffers a psychological cost, that is, feels guilty, if he lets down others' expectations. Correspondingly, a plethora of psy-game theory-driven experiments have focused on guilt aversion as a potential driver of pro-social behavior in social dilemma games (see the survey of Battigalli and Dufwenberg, 2022). The overwhelming majority of these experiments are based on two social dilemma games, the dictator and the trust games.[2] A common feature of these two games is that co-players are vulnerable, that is, either their final payoff or the use of their initial endowment depends on the actions of the decision-maker. However, is co-players' vulnerability a necessary condition to induce guilt or is guilt only driven by the willingness to not disappoint the partner's expectations even if this partner is not vulnerable? Does the nature of vulnerability matter in guilt induction? Does the partner's responsibility modulate the importance of vulnerability in inducing guilt? The present study addresses these three questions.

The previous literature on guilt aversion has not explored these questions, although experimental tests have shown that guilt aversion is modulated by a series of factors. They have focused on the role of the communication of others' expectations and on the very nature of these expectations. They have shown that communication greatly facilitates the expression of the trustee's guilt aversion, as evidenced in the milestone paper of Charness and Dufwenberg (2006) and replicated in many experimental papers since (*e.g.*, Attanasi et al., 2013; Bracht and Regner, 2013, Kawagoe and Narita, 2014; Balafoutas and Sutter, 2017; Attanasi et al., 2019a). With respect to the nature of expectations, "reasonable" expectations appear more likely to be taken into account by guilt-averse players. Khalmetski (2016), Balafoutas and Fornwagner (2017), and Danilov et al. (2021) reported an inverse-U shaped relationship between second-order beliefs and sharing decisions: dictators are less pro-social when they deem that recipients expect to receive too little or too much. Moreover, the emergence of trustees' guilt aversion is facilitated by the perceived legitimacy of the trustor's normative expectations (Andrighetto et al., 2015; Pelligra et al., 2020).

[1]This theory departs from traditional game theory in assuming that players' utilities do not only depend on their decisions but also on their beliefs about decisions, beliefs, or information (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009).

[2]Considered together, the dictator and trust games currently represent, to the best of our knowledge, the focus of 75% of published psy-game experimental studies on guilt aversion (see Table A.1 in Appendix A).

Previous studies have considered vulnerability but not in connection with guilt aversion. In particular, the moderating role of vulnerability has been studied with respect to outcome-based preferences in the trust game (Cox et al., 2016; Engler et al., 2018b). These studies have shown that the trustees' returns increase with the vulnerability of the trustor. However, in these studies the trustors' vulnerability depends on their decision[3] and beliefs were not elicited. Hence, these studies remained silent on the impact of vulnerability on guilt aversion. For its part, the literature on guilt-aversion has never questioned the potential role of the co-players' vulnerability. In fact, Battigalli and Dufwenberg (2007) (and the following applications of their model) have only explored situations where guilt-averse behaviors are directed toward a vulnerable player. Yet, guilt could very well be induced by the betrayal of expectations even of someone whose payoffs are not exposed.[4] Our work aims to bridge the gap between these two strands of the literature.

We consider that not only the existence but also the nature of vulnerability may matter in the induction of guilt aversion. We distinguish between *ex-post* and *ex-ante* vulnerability. We define a player as *ex-post vulnerable* if her material payoff depends on the actions of the decision-maker. We define a player as *ex-ante vulnerable* if her initial endowment can be entrusted to the decision-maker. In the dictator game the recipient is ex-post vulnerable, while in the trust game the trustor is both ex-ante and ex-post vulnerable. Bellemare et al. (2017) contrasted the two games in a single study and found no difference in the intensity of guilt aversion. This suggests that the combined effect of both types of vulnerability (like in the trust game) is not additive. However, we lack a comparison with only ex-ante vulnerable co-players. A first step in this direction has been taken by Attanasi et al. (2019b) who compared guilt aversion toward ex-ante vulnerable co-players vs. ex-post vulnerable co-players. They reported no difference, be it in the proportion of guilt averse players or in the intensity of guilt aversion. Their results provide indirect evidence that none of the two types of vulnerability of co-player is a necessary condition to trigger guilt.

Altogether, no final conclusion can be drawn from these studies, as they all lack a control

[3] Cox et al. (2016) considered that the trustor is vulnerable if she made a choice such that the maximum payoff she can obtain—assuming that the trustee is selfish—is lower than the maximum payoff she could have obtained otherwise—again assuming a selfish trustee. Engler et al. (2018b) defined three degrees of vulnerability in a trust game: the trustor is either (i) not vulnerable if she made a choice such that the minimum payoff she can obtain by entrusting her endowment is higher than the payoff she could have obtained by not entrusting it; (ii) vulnerable if she made a choice such that the two payoffs she can obtain by entrusting her endowment are respectively lower and higher than the payoff she could have obtained by not entrusting it; (iii) very vulnerable if she made a choice such that the maximum payoff she can obtain by entrusting her endowment is lower than the payoff she could have obtained not entrusting it.

[4] An obvious, certainly extreme, example in real settings is the guilt experienced when not respecting the last wishes of a deceased.

condition with no vulnerability at all and they do not allow a comparison, in a single study, of all the possible combinations of ex-ante and ex-post vulnerability. With the aim of providing such a comparison, the present study builds on Attanasi et al. (2019b) and introduces four variations of a three-player Trust mini-game with a passive player (Quasi-Trust mini-games, henceforth). These variations allow us to systematically compare the four possible combinations of vulnerability: no vulnerability, ex-ante vulnerability, ex-post vulnerability, ex-ante and ex-post vulnerabilities. This design also offers the possibility to test whether the responsibility of a vulnerable player makes a difference in the willingness to avoid to disappoint her (by comparing an active player (A) whose intentions are observable and a passive player (C) with the same type of vulnerability). Finally, varying the nature of games allows us to identify the combined effect of responsibility and nature of vulnerability on inducing guilt aversion.

The four Quasi-Trust mini-games are: the Investment game, the Reversed-Investment game, the Donation game (similar to Attanasi et al., 2019b), and the Exploitation game. In each game, the second mover (B) can be entrusted by the first mover (A) with a sum of money coming from the endowment of another player (A or C, depending on the game); then, he can redistribute this money between himself and another player (A or C, depending on the game).[5] The four Quasi-Trust mini-games are highly comparable since they share: for each player, the same initial endowment; for each of the two active players, the same set of strategies; for the potentially guilt averse player B, the same material payoff given the game terminal node (and thus, the same best-reply function if he is selfish). These games differ only in which player's vulnerability and which type of vulnerability is activated: ex-ante and ex-post vulnerabilities can be activated for the same player (A or C) – leaving the second one not vulnerable – or can be distributed between players A and C (A is ex-ante and C ex-post vulnerable, or the reverse).

The decisions of player B, which are the focus of the present study, can then be contrasted across the four games (i.e., across the four combinations of vulnerability) that are played within-participants. Between-participants we manipulate whether player B's decisions are elicited conditional on the first-order beliefs of either player A (active) or player C (passive). Therefore, we have a 4x2 design, which allows us to test the (in)dependence of guilt aversion from the co-players' vulnerability and responsibility.

From a theoretical point of view, we rely on a portable model of lexicographic altruism

[5] In each game, players A and C are denoted as female ("she") and player B denoted as male ("he").

and role-dependent guilt which provides predictions for the entire set of games. We assume that both players A and B can be altruistic toward the most disadvantaged player, while only player B can feel guilty. This model concludes that the partner's vulnerability has no effect on triggering guilt: player B may feel guilty even if the partner is not vulnerable, even if he cannot observe the partner's intentions (such as when the latter is a simple observer), and even toward a co-player who expresses bad intentions toward a passive player (such as in the Exploitation game). Guilt sensitivity is mainly triggered by the role in the game.

Our experimental results reveal no significant difference in the proportion of guilt-averse B-subjects across our Quasi-Trust mini-games, with a relevant fraction of B-subjects expressing guilt aversion even toward a player who is not vulnerable. This lack of significant difference suggests that vulnerability, its nature, and its link with responsibility and good or bad intentions do not modulate the trustee's guilt aversion in a Quasi-Trust game. We interpret such insensitivity of guilt aversion to the co-player's vulnerability as further support to guilt mainly being role-dependent in two-stage games with asymmetric roles, as previously suggested by Attanasi et al., 2016.

The remainder of the paper is organized as follows. Section 2 presents our four new games and their rationale given our empirical interest in the impact of the partners' vulnerability. Section 3 introduces our theoretical model and related predictions. Section 4 describes the experimental design. Section 5 presents the experimental results and Section 6 concludes.

2 The Quasi-Trust Mini-Games

To manipulate vulnerability, we introduce four Quasi-Trust games with three players: the Investment game (Figure 1), the Reversed-Investment game (Figure 2), the Donation game (Figure 3) and the Exploitation game (Figure 4). In each game, players A and B are active while player C is passive. Figures 1-4 display material payoffs according to the players' alphabetical order.

Each game unfolds as follows. A is the first mover, she can choose *In* or *Out*. If A chooses *Out*, the game ends with material payoffs corresponding to the players' initial endowments (170 ECU for A, 100 ECU for B, 30 ECU for C).[6] If A chooses *In*, she sends 25 ECU to B, with this amount being taken either from player A's or from player C's endowment (corresponding to ex-ante vulnerability), depending on the game. After *In*, player B decides how to allocate

[6] All material payoffs are expressed in Experimental Currency Units (ECU) where 10 ECU = €1 (see the experimental procedures in Section 4.3).

the 25 ECU between himself and another player, this player being A or C (corresponding to ex-post vulnerability), depending on the game. In particular, if B chooses *Left*, he transfers 5 ECU to another player and keeps 20 ECU for himself; if B chooses *Right*, he transfers the 25 ECU to this other player. Each ECU transferred by B to another player is doubled, which captures the positive externality of trust.[7]

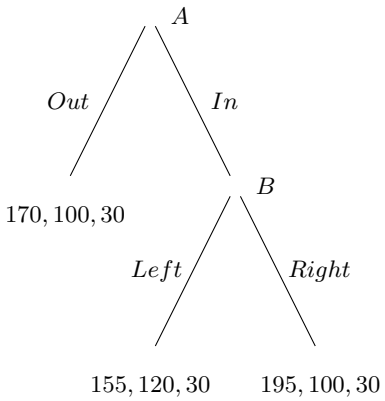

Figure 1: The Investment Game

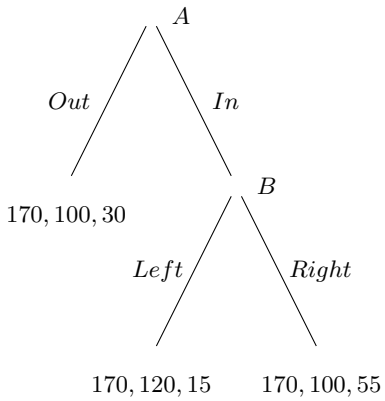

Figure 2: The Reversed-Investment Game

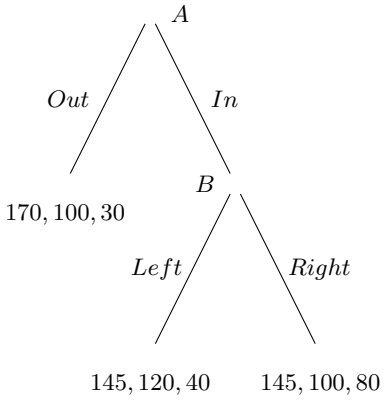

Figure 3: The Donation Game

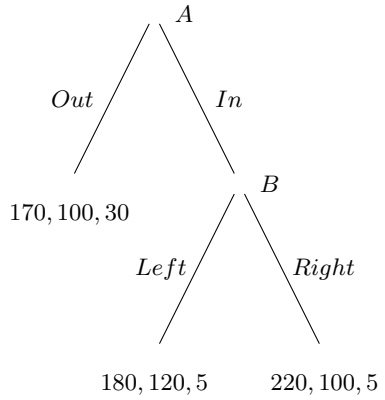

Figure 4: The Exploitation Game

In the **Investment game** (Figure 1), A can entrust B with 25 ECU taken from her own endowment. B then decides how to allocate these 25 ECU between A and himself. In this game, B's choice affects both the use of A's initial endowment and A's material payoff (i.e., A is both ex-ante and ex-post vulnerable), but it does not affect C (i.e., C is not vulnerable).

[7]Several game-independent features of the final distributions of material payoffs are worth noting. First, given the terminal node, B's material payoff is the same across the four games: if B chooses *Right* after *In*, his material payoff corresponds to his initial endowment (*Out*); if B chooses *Left* after *In*, his material payoff corresponds to his initial endowment plus the 20 ECU that he takes for himself. However, the payoff manipulation across the four games affects A's and C's payoffs (see Figure 1 to Figure 4). Next, no decision can lead to the equalization of payoffs between two or three players. Hence, no payoff distribution should be more salient than others. Furthermore, the ranking of payoffs cannot be affected by the players' decisions, which limits social comparison motives in decision making. Finally, the total surplus at a given terminal node is the same across games, this way keeping efficiency concerns constant across games.

The Investment game is a simplified version (mini-game) of the classical Trust game (see Berg et al., 1995; Buskens and Raub, 2013; Attanasi et al., 2016), with the additional feature of an external observer, C, whose payoff is affected neither by A's (trustor), nor B's (trustee) actions.

In the **Reversed-Investment game** (Figure 2), A can entrust B with 25 ECU taken from C's endowment. B then decides how to allocate these 25 ECU between C and himself. In this game, B's choice affects both the use of C's initial endowment and C's material payoff but it does not affect A, that is, A is not vulnerable and C is both ex-ante and ex-post vulnerable. Thus, the Reversed-Investment game is a modified version of the Investment game where all monetary consequences of A's investment choice fall on C: A invests C's endowment and the doubled amount can enrich C.

In the **Donation game** (Figure 3), A can entrust B with 25 ECU taken from her own endowment. B then decides how to allocate these 25 ECU between C and himself. In this game, B's choice affects both the use of A's initial endowment and C's material payoff, that is, A is ex-ante vulnerable and C is ex-post vulnerable. Thus, the Donation game is a modified version of the Investment game where the positive monetary consequences of A's investment choice fall on C: A invests her endowment and the doubled amount can enrich C. This is similar to the Embezzlement game of Attanasi et al. (2019b).

In the **Exploitation game** (Figure 4), A can entrust B with 25 ECU taken from C's endowment. B then decides how to allocate these 25 ECU between A and himself. In this game, B's choice affects both the use of C's initial endowment and A's material payoff, that is, A is ex-post vulnerable and C is ex-ante vulnerable. Thus, the Exploitation game is a modified version of the Investment game in which the negative monetary consequences of A's investment choice fall on C: A invests C's endowment and the doubled amount can enrich A.

Figure 5 summarizes the manipulation of A's and C's vulnerability across the four games.

3 Theoretical Model and Hypotheses

In this section, we develop a theoretical model of lexicographic altruism and role-dependent guilt based on the work of Attanasi et al. (2019b). After describing the players' utility functions, we analyze A's and B's best-reply functions. Finally, we elaborate theory-driven hypotheses on A's and B's behavior. We denote player j's material payoff as π_j, with $j \in \{A, B, C\}$, at each terminal node $z \in \{O, L, R\}$ of the games, that is, respectively, for each

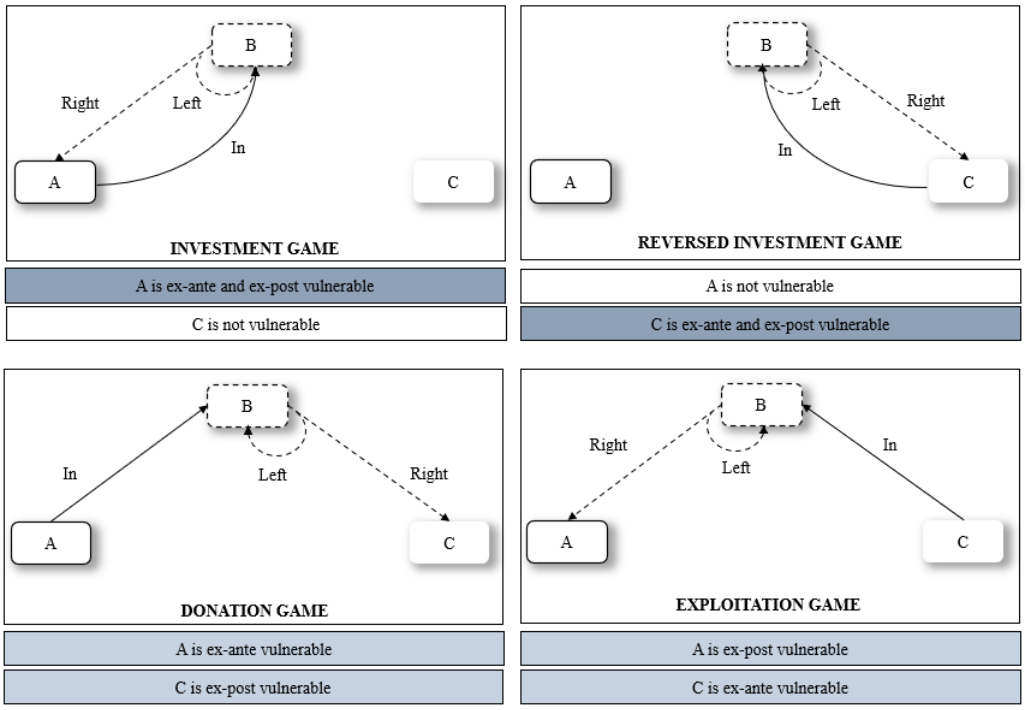

Figure 5: Vulnerability in the four Quasi-Trust mini-games

Notes: In each panel, plain lines indicate which player endowment is used by A to transfer money to B through strategy *In*; dashed lines indicate player B's strategies. If B chooses *Left*, only 5 out of 25 ECU are transferred to another player, and the rest is kept by B. If B chooses *Right*, all the 25 ECU are transferred to another player, generating higher positive externalities since each ECU transferred by B is doubled. Below each game, boxes indicate the vulnerability of A and C (white for no vulnerability, light grey for vulnerability in one dimension, dark grey for vulnerability in both dimensions

.

terminal history *Out*, *Left* after *In*, and *Right* after *In*.

3.1 Utility Functions

Since C is passive, we assume that she is purely self-interested. Therefore, **C's utility function** coincides with her material payoff, that is, $U_C(z) = \pi_C(z)$ for each $z \in \{O, L, R\}$. This assumption is motivated by the fact that, in each game and for each terminal history, C always gets the lowest payoff in the triplet.

As for **A's utility function**, we assume that she can be altruistic toward both B and C, since at each terminal node z, A always gets the highest payoff independently from the game and the strategy profile in that game. We also assume that A's altruistic preferences toward disadvantaged players are lexicographic. Precisely, since C is always the most disadvantaged player and B is always the second most disadvantaged player, A is altruistic only toward C when C's payoff depends on A's strategy, and only toward B when B's payoff depends on

A's strategy but C's payoff does not.[8] Therefore, A can be altruistic toward player C in the Reversed-Investment, Donation and Exploitation games, and she can be altruistic toward B in the Investment game.

We model A's feeling of altruism toward player $h \in \{B, C\}$, F_{Ah}, as A's utility derived from the payoff of h. It is the product of two terms: $\phi_{Ah} \geq 0$, A's sensitivity to altruism toward h, and $\pi_h(z)$, h's material payoff. With this, A's utility (Eq. (1)) is composed of her material payoff and her feeling of altruism toward $h \in \{B, C\}$ (Eq. (1)):

$$U_A(\phi_{Ah}, z) = \pi_A(z) + F_{Ah}(\phi_{Ah}, z), \text{where } F_{Ah}(\phi_{Ah}, z) = \phi_{Ah} \cdot \pi_h(z) \tag{1}$$

with $h = B$ in the Investment game and $h = C$ in the remaining three games.

Let us now introduce **B's utility function**. Besides B's concern for his own payoff, π_B, we assume that B has lexicographic altruistic preferences toward disadvantaged players modeled like those of player A (see F_{Bh}, namely B's utility derived from the payoff of player $h \in \{A, C\}$, in Eq. (3)). By construction of the four games, since C is always the most disadvantaged player and A is always the most advantaged player, B can be altruistic only toward C. He is altruistic toward C when C's payoff depends on his strategy, and toward no player when C's payoff does not depend on his strategy (in the latter case, there is no player more disadvantaged than him whose payoff he can increase). Therefore, F_{Bh} in Eq. (3) essentially coincides with F_{BC} in each of the four games ($h = C$). The latter only has a strategic impact in the Reversed-Investment and Donation games in which C's material payoff depends on B's strategy: from Eq. (3), $F_{BC}(\phi_{BC}, R) > F_{BC}(\phi_{BC}, L)$, that is, *Right* after *In* is a more altruistic strategy than *Left* after *In*. In the Investment and Exploitation games, in which C's material payoff does not depend on B's strategy, it is $F_{BC}(\phi_{BC}, R) = F_{BC}(\phi_{BC}, L)$, hence B's altruism is irrelevant.

Furthermore, in line with the role-dependent guilt model of Attanasi et al. (2016), we assume that B can feel guilty due to his role in the game, whereas A does not.[9] B's feeling of guilt, G_{Bjk}, with $j, k \in \{A, C\}$ in Eq. (4), represents his disutility derived from letting down j's beliefs on the strategy he will select, which will affect k's payoff, with j not necessarily equal to k. More precisely, it is the product of two terms: $\gamma_{Bjk} \geq 0$, B's guilt sensitivity

[8] The assumption of lexicographic altruistic preferences is broadly consistent with inequity-aversion models (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), since C is the most disadvantaged player.

[9] See the discussion in Attanasi et al. (2016), p. 649, where they argue that role dependence of guilt preferences is plausible in asymmetric games (see, *e.g.*, Attanasi et al., 2013, 2019a, for indirect experimental evidence corroborating this assumption). In particular, they discuss how the assumption that sensitivity to guilt is triggered only when playing in the role of trustee (and not in the role of trustor) in the trust game resonates with the evolutionary psychology of emotions and the conceptual act theory of emotion. Similar arguments can be provided in support of sensitivity to guilt being triggered only when playing in the role of player B (and not in the role of player A) in our four Quasi-Trust mini-games of Figure 1 to Figure 4.

about j's beliefs when B's strategy affects k's payoff; and the difference, if positive, between j's beliefs about k's payoff after In, $\mathbb{E}_j[\pi_k(z|In)]$, and k's actual material payoff after In, $\pi_k(z|In)$. More precisely, if $\mathbb{E}_j[\pi_k(z|In)] = \alpha_{jB} \cdot \pi_k(R) + (1 - \alpha_{jB}) \cdot \pi_k(L) > \pi_k(z|In)$ (where α_{jB} is j's first-order belief that B chooses *Right* after In), then B feels guilty from letting down j's beliefs on k's payoff; otherwise, his guilt feeling G_{Bjk} is null since he does not let down j's beliefs on k's payoff. Eq. (2) expresses B's utility after In, with altruism and guilt feelings represented in respectively Eq. (3) and Eq. (4) for $j, k \in \{A, C\}$:

$$U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, z|In) = \pi_B(z|In) + F_{BC}(\phi_{BC}, z|In) - G_{Bjk}(\gamma_{Bjk}, \alpha_{jB}, z|In) \quad (2)$$

$$\text{where } F_{BC}(\phi_{BC}, z|In) = \phi_{BC} \cdot \pi_C(z|In) \quad (3)$$

$$\text{and } G_{Bjk}(\gamma_{Bjk}, \alpha_{jB}, z|In) = \gamma_{Bjk} \cdot \max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\} \quad (4)$$

We anticipate here that we implement an experimental design in which the impact of guilt sensitivity toward A can be analyzed separately from the impact of sensitivity toward C (see Section 4). In fact, we use a between-subject design to elicit B's belief-dependent strategy conditional on either A's (treatment A) or C's (treatment C) first-order beliefs about *Right* if *In*. Thus, as for Eq. (4), we elicit guilt sensitivity γ_{Bjk}, with $j = A$ in treatment A and $j = C$ in treatment C regardless of the Quasi-Trust mini-game. In treatment A, in which B's strategy is elicited conditional on A's first-order beliefs ($j = A$, hence G_{BAk}), the standard Battigalli and Dufwenberg (2007) definition of guilt aversion ($k = A$, hence G_{BAA}) only applies in the Investment and Exploitation games, while the extended definition ($k = C$, hence G_{BAC}) also applies in the Reversed-Investment and Donation games. Correspondingly, in treatment C in which B's strategy is elicited conditional on C's first-order beliefs ($j = C$, hence G_{BCk}), the standard Battigalli and Dufwenberg (2007) definition of guilt aversion ($k = C$, hence G_{BCC}) only applies to the Reversed-Investment and Donation games, and the extended definition ($k = A$, hence G_{BCA}) also applies to the Investment and Exploitation games.

3.2 Best-Reply Analysis and Hypotheses

We elaborate our hypotheses relying on best-reply analysis rather than on Bayesian equilibrium. Indeed, a standard equilibrium analysis has no compelling foundation for games played one-shot, like ours, and in experiments on other-regarding preferences (see Section 6.2 of Attanasi et al., 2016).

3.2.1 Player A's Best-Reply Functions and Hypotheses

As we do not use Bayesian equilibrium as a solution concept and because we are mainly interested in B's behavior, here we only present a brief summary of A's best-reply analysis (the full analysis can be found in Appendix B). The aim of this section is to show that given a (type, belief) pair of player A, her predicted behavior is game-dependent.

The strategy *In* is the altruistic one in all games but the Exploitation game, in which *Out* is the altruistic strategy (see Figures 1-4). In the **Investment Game**, a selfish or lightly-altruistic A chooses In more, the higher is her first-order belief α_{AB} that B chooses *Right* after *In*; a highly-altruistic A chooses In regardless of α_{AB}. Thus, given α_{AB}, the higher ϕ_{AB}, the higher the likelihood that A chooses In. In the **Reversed-Investment Game**, a selfish A chooses In regardless of α_{AB}; any altruistic player A chooses In only for high enough α_{AB}. In the **Donation Game**, a selfish or lightly-altruistic A never chooses In regardless of α_{AB}; a highly-altruistic A chooses In only for high enough α_{AB}. In the **Exploitation Game**, a selfish or lightly-altruistic A always chooses In regardless of α_{AB}; a highly-altruistic A chooses In only for low enough α_{AB}.

Therefore, A's altruism leads to belief-dependent behavior. Furthermore, and more importantly, A's belief-dependent behavior is also game-dependent. In fact, the relationship between altruism sensitivity ϕ_{Ah} and the first-order belief α_{AB} that leads to In as a best-reply strategy differs across the four Quasi-Trust mini-games. Recall that in our experiment A-subjects are unaware of the treatment when they make their choices, hence A's behavior should be treatment-independent.

Our hypotheses refer to two aspects of A's choices. First, H.A. 1 and H.A. 2 address A's lexicographic altruism in each game taken separately: as the theoretical predictions in Table B.1 in Appendix B shows, a more trustful A-player is more willing to choose In regardless of the game, while the interplay between altruism sensitivity and willingness to choose In depends on the game. As for the latter, H.A. 3 and H.A. 4 specify A's motivation behind In choices across games. If these four hypotheses are supported, this would suggest that A's intention behind In is to increase C's payoff in the Reversed-Investment and Donation games, while she wishes to increase her own payoff in the Investment and Exploitation games.

H.A. 1. [*Choice-belief correlation*] The frequency of *In* choices by A-subjects increases in their first-order belief about B-subjects choosing *Right* in each game.

H.A. 2. [*Choice-type correlation*] The frequency of *In* choices by A-subjects increases in their

sensitivity to altruism in the Investment and the Donation games. It decreases in A-subjects' sensitivity to altruism in the Reversed-Investment and the Exploitation games.

H.A. 3. [*Choice under beliefs of a distrustful A*] For A-subjects thinking that *Left* is the most likely action of B-subjects, the frequency of In choices in the Donation game is lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and lightly-altruistic types; (iii) in the Investment game for highly-altruistic types.

H.A. 4. [*Choice under beliefs of a trustful A*] For A-subjects thinking that *Right* is the most likely action of B-subjects, the frequency of In choices in the Investment game is: (i) the same as in the Reversed-Investment game, regardless of the altruistic type; (ii) higher than in the Donation game for selfish and lightly-altruistic types; (iii) higher than in the Exploitation game for highly-altruistic types.

3.2.2 Player B's Best-Reply Functions and Hypotheses

In each of the four Quasi-Trust mini-games, if B chooses *Right* after *In*, he entirely transfers to another player the amount of money that A's *In* choice has entitled him to manage. If instead he chooses *Left* after *In*, he only transfers a small portion (20%) of that amount. Therefore, relying on Eqs. (2–4), for each treatment (A and C) we define B's *Willingness-to-Transfer function* (WT) as the difference between his utility from playing *Right* after *In* and his utility from playing *Left* after *In*. Both terms are expected utilities since B forms beliefs about the first-order beliefs α_{jB} of the co-player j toward whom he may feel guilty ($j = A$ in treatment A and $j = C$ in treatment C).[10] These are his conditional second-order beliefs $\beta_{Bj} = \mathbb{E}_B[\alpha_{jB}|In]$ for $j \in \{A, C\}$, that is, conditional on A choosing *In*.[11] The higher B's willingness to transfer the money that A's *In* choice has entitled him to manage, the more player B prefers to choose *Right* rather than *Left*:

$$\begin{aligned} WT(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, z|In) &= \mathbb{E}_B[U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, R)] - \mathbb{E}_B[U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, L)] \\ &= \pi_B(R) - \pi_B(L) + \phi_{BC} \cdot [\pi_C(R) - \pi_C(L)] + \\ &\quad \gamma_{Bjk} \cdot \beta_{Bj} \cdot [\pi_k(R) - \pi_k(L)] \end{aligned} \tag{5}$$

More precisely, B chooses *Right* after In if $WT > 0$ in Eq. (5), and *Left* otherwise. Note that given their common structure, in each of the four games it is $\pi_B(R) - \pi_B(L) = -20$. With

[10] In each game, we assume that B best-responds *as if* he had truly observed A's move. This holds by standard expected-utility maximization, except for the cases where B is certain that A has chosen *Out*. Thus, we need the additional assumption that B has a belief conditional on *In*, even if he is certain of *Out*. Indeed, in our experiment (see Section 4) B's decision is made under the strategy method, *i.e.*, both when A has chosen *Out* and when she has chosen *In*.

[11] More precisely, we reason as if B has a point belief β_{Bj} about α_{jB} conditional on *In*.

this, we can find player B's best-reply strategy as a function of his sensitivity to altruism toward player C, ϕ_{BC}, his second-order belief β_{Bj} that he will choose *Right* after *In*, and his sensitivity γ_{Bjk} to guilt toward the player j on whom B's second-order belief β_{Bj} relies.

In the **Investment Game**, B's strategy does not affect C's payoff, thus B cannot be altruistic toward C (by construction, $F_{BC}(\phi_{BC}, L) = F_{BC}(\phi_{BC}, R)$ in Eq. (3)). Furthermore, B's strategy affects A's payoff, hence $k = A$ in Eq. (5). WT in Eq. (5) reduces to:

$$\pi_B(R) - \pi_B(L) + \gamma_{BjA} \cdot \beta_{Bj} \cdot [\pi_A(R) - \pi_A(L)] \tag{6}$$

By substituting the game payoffs of Figure 1, Eq. (6) becomes $-20 + 40 \cdot \gamma_{BjA} \cdot \beta_{Bj}$, which is strictly positive for all type-belief pairs (γ_{BjA}, β_{Bj}) such that $\gamma_{BjA} \cdot \beta_{Bj} > 1/2$. Therefore, a guilt-averse B is more willing to choose *Right* after *In* for higher guilt sensitivity γ_{BjA} and higher second-order belief β_{Bj} of *Right* after *In*. This relationship holds both in treatment A, that is, for type-belief pairs (γ_{BAA}, β_{BA}), and in treatment C, that is, for (γ_{BCA}, β_{BC}).

In the **Reversed-Investment Game**, B's strategy affects C's payoff hence $k = C$ in Eq. (5), which becomes:

$$\pi_B(R) - \pi_B(L) + (\phi_{BC} + \gamma_{BjC} \cdot \beta_{Bj}) \cdot [\pi_C(R) - \pi_C(L)] \tag{7}$$

By substituting the game payoffs of Figure 2, Eq. (7) becomes $-20 + 40 \cdot (\phi_{BC} + \gamma_{BjC} \cdot \beta_{Bj})$, which is strictly positive for all type-belief pairs ((ϕ_{BC}, γ_{BjC}), β_{Bj}) such that $\phi_{BC} + \gamma_{BjC} \cdot \beta_{Bj} > 1/2$. Therefore, a guilt-averse B is more willing to choose *Right* after *In* for higher guilt sensitivity γ_{BjC} and higher conditional second-order belief β_{Bj} of *Right* after *In*. This relationship holds both in treatment A, that is, for type-belief pairs (γ_{BAC}, β_{BA}), and in treatment C, that is, for (γ_{BCC}, β_{BC}). Furthermore, independently from the treatment, the higher ϕ_{BC}, B's sensitivity to altruism toward C, the lower both the guilt sensitivity γ_{BjC} and the second-order belief β_{Bj} required for B to choose *Right* after *In*. Finally, for high enough sensitivity to altruism (i.e., $\phi_{BC} > 1/2$), player B chooses *Right* after *In*, regardless of his second-order belief β_{Bj}.

In the **Donation Game**, B's strategy affects C's payoff, hence $k = C$ in Eq. (5). Thus, WT in this game is the same as in Eq. (7). By substituting the game payoffs of Figure 3, given the similar structure between the Reversed-Investment and the Donation games ($\pi_B(R) - \pi_B(L) = -20$ and $\pi_C(R) - \pi_C(L) = 40$), we find the same subset of type-belief pairs ((ϕ_{BC}, γ_{BjC}), β_{Bj}) for which Eq. (7) is strictly positive, that is,, $\phi_{BC} + \gamma_{BjC} \cdot \beta_{Bj} > 1/2$.

Therefore, independently from treatment, the same considerations made for the Reversed-Investment game hold in the Donation game.

Finally, in the **Exploitation Game**, B's strategy does not affect C's payoff, thus B cannot be altruistic toward C ($F_{BC} = 0$). In contrast, B's strategy affects A's payoff, hence $k = A$ in Eq. (5). Thus, WT in this game is the same as in Eq. (6). By substituting the game payoffs of Figure 4, given the similar structure between the Investment and the Exploitation games ($\pi_B(R) - \pi_B(L) = -20$ and $\pi_A(R) - \pi_A(L) = 40$), we find the same subset of type-belief pairs (γ_{BjA}, β_{Bj}) for which Eq. (6) is strictly positive, that is, $\gamma_{BjA} \cdot \beta_{Bj} > 1/2$. Independently from the treatment, the same considerations made for the Investment game hold in the Exploitation game.

We have two families of hypotheses for B-subjects: a first one considering, for each game taken separately, the correlations between B's choices and his second-order belief (H.B. 1) or type (H.B. 2); and a second one comparing B's decisions across games (H.B. 3 to H.B. 5).

Based on this best-reply analysis, we can derive our hypotheses about B-subjects' behavior. Taken together, H.B. 1 and H.B. 2 postulate that guilt is activated in each of the eight game-treatment combinations. These hypotheses are at the core of our extension of Battigalli and Dufwenberg (2007). They contrast with the predictions from Battigalli and Dufwenberg (2007) and follow-up studies that expect guilt to arise in only four treatment-game combinations (those in which B's strategy and second-order beliefs are conditioned to the first-order beliefs of a player whose payoff depends on B's strategy, that is, the Investment and Exploitation games in treatment A and the Reversed-Investment and Donation games in treatment C).[12]

H.B. 1. [*Choice-belief correlation*] The frequency of *Right* choices by B-subjects increases in their second-order beliefs about *Right* in each of the four games.

H.B. 2. [*Choice-type correlation*] Given a positive second-order belief, the frequency of *Right* choices of B-subjects increases with: (i) their altruism sensitivity only in the Reversed-Investment and Donation games; (ii) their guilt sensitivity in each of the four games.

After assuming that B can feel guilty also when disappointing the beliefs of a player whose payoff is not affected by B's decision, we now examine the frequency of such guilt-averse behavior. Since our model is silent on this issue, we rely on Attanasi et al. (2019b)

[12]Note that, given A's first-order belief of *Right*, the same *In* choice in different games would signal an A's different sensitivity to altruism. Therefore, if B cares about the different intentions behind A's *In* choice, B's guilt sensitivity should also be game-dependent. However, our study relies on the opposite intuition that B's belief-dependent behavior is game-independent.

who tested this hypothesis in the Donation game and detected no significant difference in B's guilt between $j = k = C$ and $A = j \neq k = C$ in Eq. (2). Therefore, H.B. 3 and H.B. 4 posit the same fraction of guilt-averse B-players across the four games and the two treatments.

H.B. 3. [*Within-subject game-independent guilt*] Within a treatment, the fraction of guilt-averse B-subjects does not differ across the four games.

H.B. 4. [*Between-subject treatment-independent guilt*] Within a game, the fraction of guilt-averse B-subjects is not significantly different across treatments.

The joint test of H.B. 3 and H.B. 4 is the most original contribution of our study. This test helps us assessing whether (i) disappointing an ex-ante vulnerable player leads to higher guilt than disappointing a non-vulnerable one; (ii) disappointing an ex-post vulnerable player leads to higher guilt than disappointing a non-vulnerable one; (iii) disappointing an ex-post vulnerable player leads to higher or lower guilt than disappointing an ex-ante vulnerable one; (iv) disappointing an ex-ante and ex-post vulnerable player leads to higher guilt than disappointing a player vulnerable on just one of these two dimensions.

Finally, relying on the assumption of lexicographic altruism, H.B. 5 asserts that B's altruism is only activated in the Reversed-Investment and Donation games. Since guilt aversion and altruism are the only other-regarding motivations of B in Eq. (2), this would ultimately lead to a smaller fraction of selfish B-subjects (i.e., always choosing *Left*) in these games.

H.B. 5. [*Game-dependent altruism*] The fraction of B-subjects who behave selfishly is significantly higher in the Investment and Exploitation games than in the Reversed-Investment and Donation games. This holds independently of the treatment.

4 Experimental Design and Procedures

In our experimental design, each subject went through the four Quasi-Trust games of Figures 1-4: Donation, Investment, Reversed-Investment and Exploitation. The games were renamed with neutral labels ("North", "South", "East", and "West"). In each game, subjects played in groups of three, with roles (A, B and C) assigned at the beginning of the session and maintained fixed across games. Groups were re-matched across games according to a perfect-stranger protocol. We randomized within-subjects the order of presentation of the four games across experimental sessions, and we varied between-subjects the treatments A and C.

4.1 Decisions and Elicitation of Beliefs

First-order belief elicitation We elicited, for each game, B-subjects' and C-subjects' first-order beliefs on the frequency of A-subjects choosing *In*. They had to report, for each game, their belief about the number of A-subjects, out of three A-subjects randomly selected in the session, who chose *In*, from 0 to 3 inclusive. We also elicited, for each game, A-subjects' and C-subjects' first-order beliefs on the frequency of B-subjects choosing *Right* after *In*. Similarly, they had to report, for each game, their belief about the number of B-subjects, out of three B-subjects randomly selected in the session, who chose *Right* conditional on A-subject choosing *In*, from 0 to 3 inclusive. For each role, one belief out of the four elicited in the four games was randomly selected at the end of the session and paid €1 if accurate.

A-subject's decision For each game, A-subjects chose between *In* or *Out*. At the end of the session, one of the four games was randomly selected to be payoff-relevant.

B-subject's decision B-subjects decided under the veil of ignorance, that is, assuming that their matched A-subject had chosen *In*. For each game, in treatment A (respectively, treatment C) B-subjects made four decisions corresponding to their matched A-subject's (respectively, C-subject's) four possible first-order beliefs on the frequency of *Right* choices conditional on *In*. In other words, in treatment A (respectively, treatment C), B-subjects could condition their decision to the possible first-order beliefs of their matched A-subject (respectively, C-subject). Given that the A-subject had chosen *In* in the game randomly selected to be payoff-relevant, the program implemented the B-subject's decision corresponding to the actual first-order belief of the A-subject (respectively, C-subject) in treatment A (respectively, treatment C). To facilitate decision making, the four possible first-order beliefs were presented in a fixed increasing order. This elicitation of decisions conditional on another player's first-order belief corresponds to the menu method of Khalmetski et al. (2015), which allows the experimenter to artificially induce second-order beliefs.[13]

Second-order belief elicitation We elicited, for each game, A-subjects' second-order beliefs on the frequency of A-subjects choosing *In*, according to their matched B-subject's and

[13]The use of the menu method is now frequent in the experimental literature on guilt aversion (Khalmetski et al., 2015; Hauge, 2016; Balafoutas and Fornwagner, 2017; Bellemare et al., 2017; Dhami et al., 2019; Bellemare et al., 2018). Although one might argue that this method elicits "cold" responses, it offers several advantages. It allows to rule out potential false-consensus effects without raising the issue of strategic reporting and without using deception. The false-consensus effect could be avoided by communicating the A-subject's (C-subject's) true beliefs to B-subjects. However, it requires choosing between two evils: if A-subjects (C-subjects) know that their beliefs will be communicated, they are likely to distort them; and if they do not know that their beliefs will be communicated, the design is arguably deceptive. The menu method avoids these drawbacks. Moreover, it allows us to study guilt aversion at the individual level and, hence, to unveil inter-individual differences that are hidden at the aggregate level (Khalmetski et al., 2015).

C-subject's in the game. In other words, A had to guess B's and C's first-order beliefs on the frequency of A-subjects choosing *In*. We also elicited, for each game, B-subjects' second-order belief on the frequency of B-subjects choosing *Right* after *In*, according to their matched A-subject and C-subject. Relying on previously elicited first-order beliefs, second-order beliefs were also elicited through asking subjects to report a number from 0 to 3, inclusive. For each role, one belief out of the four elicited (one for each of the four games) was randomly selected at the end of the session and paid €1 if accurate.

4.2 Elicitation of Individual Preferences

In the second part of the experiment we elicited social preferences via the Social Value Orientation (SVO) test (Murphy et al., 2011). In the role of a decision maker, subjects made fifteen allocation choices between a decision maker and a passive player. They were paid for two randomly selected periods: one as a decision maker, one as a passive player.

Additionally, at the end of the session we collected non-incentivized measures of individual preferences, using the Guilt and Shame Proneness (GASP) questionnaire (Cohen et al., 2011). Moreover, subjects had to self-report their attitudes toward risk, patience and guilt proneness.[14] Finally, we collected socio-demographic characteristics, including gender, age, major and number of past participation in economic experiments.

4.3 Procedures

The experiment was conducted at GATE-Lab, Lyon, France. It was computerized using z-Tree (Fischbacher, 2007). Subjects were recruited mainly from the undergraduate student population of local business, engineering and medical schools, using Hroot (Bock et al., 2014). 288 subjects participated in a total of 17 sessions. 57% were female and the average age was 22 years. Table D.1 in Appendix D shows that the mean individual characteristics are similar across treatments.

The session consisted of two parts. The instructions (see Appendix C) for the first part were distributed before each stage. The first stage described the four games. The experi-

[14]Risk aversion and patience were measured by the following questions: "Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" (Dohmen et al., 2011), and "Are you generally an impatient person, or someone who always shows great patience?" (Vischer et al., 2013). We adapted Moulton et al. (1966) to phrase the question on guilt proneness in a similar manner as for risk aversion and patience: "Are you generally a person who easily feels guilty or is it difficult to make you feel guilty?". Subjects rated how "easy" it is to make them feel guilty on a scale from 0 to 10, that is, with the same rating scale used to answer the two questions on how "willing to take risk" and how "patient" they are.

menter made sure that all subjects had completed correctly a comprehension questionnaire before moving on to the second stage. At the beginning of the second stage, subjects were informed of their role. Then, we elicited the subjects' first-order beliefs and A-subjects made their decisions. In the third stage, B-subjects made their decisions. Meanwhile, A- and C-subjects could solve sudoku-puzzles to avoid that their immediate neighbors in the lab could identify their role. In the fourth stage, we elicited the A- and B-subjects' second-order beliefs while C-subjects could solve sudoku puzzles. In the second part of the experiment, we implemented the SVO test and the questionnaires on risk and guilt proneness. Then, subjects received feedback on their payoff and the decisions that were payoff-relevant, and they finally completed the socio-demographic questionnaire.

Each session lasted about 75 minutes. Game payoffs were expressed in Experimental Currency Units (ECU) with 10 ECU = €1. Average earnings were €17 (S.D. = 5.91), including payment for accurate beliefs and a €5 show-up fee.

5 Results

We first briefly summarize the main findings regarding A-subjects' behavior, with all details given in Appendix D.2. Then, we focus on B-subjects.

5.1 A-Subjects' Behavior

As expected, the choice of *In* by the 96 A-subjects varies considerably across games. Pooling the two treatments, *In* is chosen by 20.83% of the A-subjects in the Donation game, 48.87 % in the Investment game, 70.83% in the Reversed-Investment game, and 75% in the Exploitation game (see also Table D.2 in Appendix D.2 that reports the proportion of *In* choices across games, by first-order beliefs). We reject the null hypothesis that the proportion of A-subjects choosing *In* is the same across games (Cochran Q test; $p = 0.000$). Consistently, pairwise comparisons show that this proportion is significantly different across games (McNemar tests; highest $p = 0.001$ for Investment vs. Reversed-Investment), except when we compare the Reversed-Investment and the Exploitation games (70.83% vs. 75.00%; $p = 0.584$).[15]

To test H.A. 1 and H.A. 2, we estimated separate Logit regressions for each game with the choice of *In* as the dependent variable, and with As' first-order beliefs and SVO angle

[15]Except when specified otherwise, the non-parametric tests are two-sided and each decision is treated as one independent observation since only one decision per participant is payoff-relevant.

as the main independent variables. The regressions and further details are reported in Table D.3 in Appendix D.2. This analysis yields the first two results about A-subjects behavior:

R.A. 1. *[Choice-belief correlation] The frequency of* In *choices by A-subjects increases significantly in their first-order belief about B-subjects choosing* Right *in the Investment and Donation games, but not in the Reversed-Investment and Exploitation games.*

R.A. 2. *[Choice-type correlation] The frequency of* In *choices by A-subjects increases significantly in their altruism sensitivity in the Investment and Donation games, but is not significantly influenced by their altruism sensitivity in the Reversed-Investment and Exploitation games.*

To test H.A. 3, we consider the choices of the A-subjects who believe that *Left* is the most likely action of the B-subjects (i.e., those with $\alpha_{AB} \leq 1/3$). We separate between selfish, lightly-altruistic and highly-altruistic A-subjects, as suggested by Table B.1 in Appendix B. We split them uniformly into these categories according to their SVO angle (further details are provided in Appendix D.2). Our analysis supports H.A. 3 and yields the following result:

R.A. 3. *[Choice under beliefs of a distrustful A] For the A-subjects who believe that* Left *is the most likely action of the B-subjects, the frequency of* In *choices in the Donation game is significantly lower than: (i) in the Reversed-Investment game for selfish types, (ii) in the Exploitation game for selfish and lightly-altruistic types, and (iii) in the Investment game for all altruistic types.*

Finally, to test H.A. 4, we consider the choices of the A-subjects who believe that *Right* is the most likely action of the B-subjects (i.e., those with $\alpha_{AB} > 2/3$). We also separate among selfish, lightly-altruistic and highly-altruistic A-subjects. Most differences between the Investment game and the other games predicted by H.A. 4 are supported by the analysis. Our last result for the A-subjects is the following:

R.A. 4. *[Choice under beliefs of a trustful A] For the A-subjects who believe that* Right *is the most likely action of the B-subjects, the frequency of* In *choices in the Investment game is: (i) not significantly different than in the Reversed-Investment game, regardless of the type; (ii) higher, although not significantly so, than in the Donation game for selfish and lightly-altruistic types; (iii) significantly higher than in the Exploitation game for the highly-altruistic type.*

5.2 B-Subjects' Behavior

Before testing our hypotheses formally, we describe B-subjects' behavior through five patterns of choices for their four induced second-order beliefs, $\beta_{Bj} \in \{0, 1/3, 2/3, 1\}$, in each game:[16]

(i) always choosing *Left*, regardless of the induced second-order beliefs, that is, choosing the payoff-maximizing (selfish) option (this represents on average 57% of the B-subjects);[17]

(ii) always choosing *Right*, regardless of the induced second-order beliefs, that is, choosing the efficiency-maximizing option (5% of the B-subjects);

(iii) switching from *Left* to *Right* as the induced second-order belief increases, that is, disclosing guilt aversion (26% of the B-subjects);

(iv) switching from *Right* to *Left* as the induced second-order belief increases (6% of the B-subjects);

(v) any other pattern of choices (6% of the B-subjects).

Figure 6 displays the distribution of B-subjects across these five patterns of choices in each game and for each treatment ($j = A$ and $j = C$) separately.[18] Note that among the five patterns of behavior identified above, our model is consistent with behaviors described in patterns (i) to (iii): they represent 87.54% of all B-subjects' behavior. More importantly, guilt-averse behavior (ii) represents 60% of all non-selfish behavior (patterns (ii) to (v)), thereby showing that guilt aversion is the prevailing social preference.

We now directly test our hypotheses. Table 1 presents the marginal effects from panel Logit regressions on the probability to choose *Right*. For each game, we report two specifications. First, we regress B-subjects' choices on their induced and their stated second-order beliefs, β_{Bj}, their altruism sensitivity (measured by their SVO angle), ϕ_{Bj}, and their self-reported guilt proneness, γ_{Bjk}. We control for the treatment (A or C) and the order in which the game was played. The second specification adds personality (risk aversion and patience) and socio-demographic controls (age, gender, frequency of past participation in experiments, majoring in business).

[16] By "induced second-order beliefs" we denote the four possible first-order beliefs of A-subjects (respectively, C-subjects) displayed on B-subjects' screens in treatment A (respectively, treatment C).

[17] The fact that the fraction of selfish B-subjects detected in our four games is on average higher than 50% is not surprising. Differently from the standard trust game, B's trustworthiness (*Right* if *In*) brings him no additional money with respect to his initial endowment, since $\pi_B(R) = \pi_B(O)$. Thus, in each game a B-subject choosing *Right* is purely driven by other-regarding preferences.

[18] In addition, Figure D.1 in Appendix D.4 analyzes the consistency of B-subjects' patterns of choices.

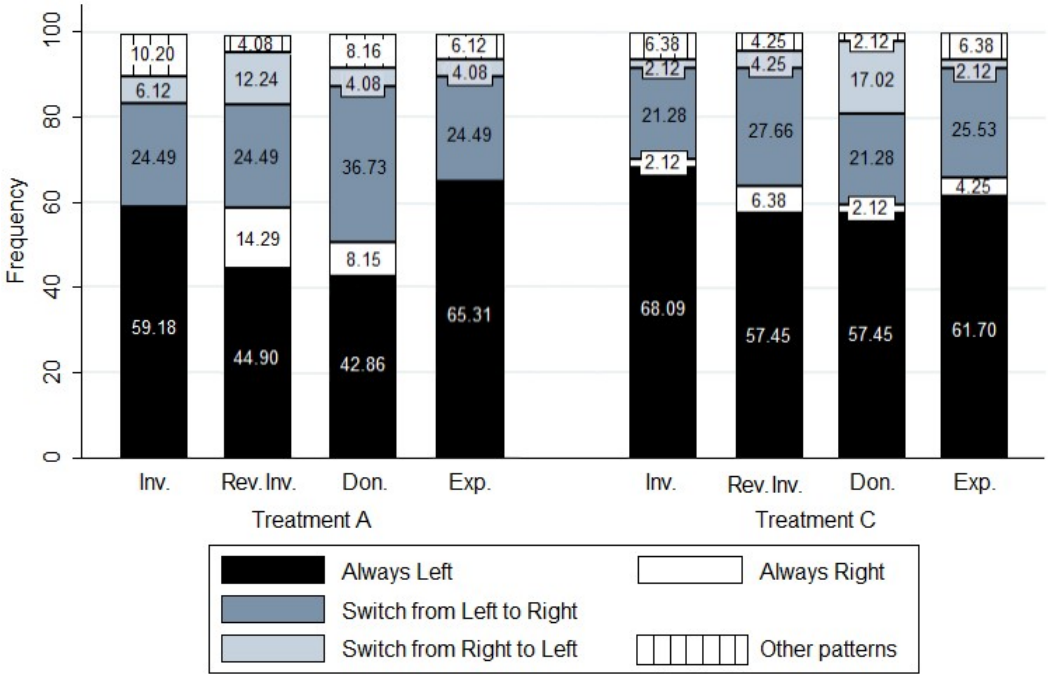

Figure 6: Distribution of B-Subjects' Pattern of Choices Across Games and Treatments

Table 1: Likelihood of B-Subjects Choosing *Right*, by Game

Game	Investment		Rev. Investment		Donation		Exploitation	
Induced SOB: β_{Bj}	0.186***	0.192***	0.187***	0.182***	0.197***	0.194***	0.208***	0.212***
	(0.045)	(0.045)	(0.048)	(0.047)	(0.050)	(0.048)	(0.045)	(0.044)
Stated SOB	0.270***	0.246***	0.434***	0.425***	0.411***	0.388***	0.245***	0.244***
	(0.075)	(0.072)	(0.087)	(0.081)	(0.096)	(0.084)	(0.064)	(0.059)
SVO Angle: ϕ_{Bj}	0.003*	0.002	0.010***	0.009***	0.004*	0.003	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Reported Guilt:	0.002	0.006	0.030***	0.033***	0.011	0.024**	0.005	0.013
γ_{Bjk}	(0.008)	(0.008)	(0.011)	(0.011)	(0.011)	(0.010)	(0.008)	(0.009)
Treatment A	0.012	0.023	0.143**	0.120*	0.168***	0.181***	-0.032	-0.004
	(0.042)	(0.045)	(0.061)	(0.061)	(0.059)	(0.054)	(0.044)	(0.046)
Order	-0.013	-0.015	-0.010	-0.002	0.007	-0.014	-0.055**	-0.057***
	(0.023)	(0.025)	(0.025)	(0.024)	(0.022)	(0.022)	(0.023)	(0.022)
Constant	-4.669**	-5.821**	-8.036***	-7.353**	-5.671***	-8.884***	-3.372	-9.407**
	(1.935)	(2.863)	(2.145)	(3.012)	(1.719)	(2.480)	(2.443)	(3.863)
Personality	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographics	No	Yes	No	Yes	No	Yes	No	Yes
N Observations	384	384	384	384	384	384	384	384
N subjects	96	96	96	96	96	96	96	96
Log-likelihood	-130.210	-126.656	-155.339	-152.365	-169.71	-159.82	-126.191	-120.278
Wald Chi2	28.82	31.69	36.24	38.12	32.30	40.39	31.78	34.36
Prob>chi2	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Table 1 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses."Induced SOB β_{Bj}" corresponds to the four β_{Bj} presented to B-subjects when making their choice of *Right* or *Left*. "Stated SOB β_{Bj}" corresponds to the second-order beliefs reported by the B-subjects in the belief elicitation stage. "Reported Guilt" takes value between 0 and 10. "SVO angle" takes value between -7.8 and 38.9. "Order" is the rank order of the game, from 1 to 4. "Personality" controls correspond to the subjects' self-reported risk aversion and patience. "Socio-Demographics" controls include age, gender, frequency of past participation in experiments, majoring in business. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

As predicted by H.B. 1, Table 1 shows that in all games, regardless of the specification, the higher is the supposed first-order belief of their matched A or C-subject (i.e., B-subjects' induced second-order belief in treatment A or C, respectively), the more likely B-subjects are to choose *Right*. The same holds for the stated second-order beliefs: the likelihood of choosing *Right* significantly increases in B-subjects' stated second-order beliefs. Both are significant at the 1% level, regardless of the specification.[19] Thus, H.B. 1 is supported in each treatment and for any measure of second-order beliefs, as stated in R.B. 1.

R.B. 1. *[Choice-belief correlation] The likelihood of B-subjects choosing* Right *significantly increases in their second-order beliefs about* Right *in each game, regardless of the treatment and for both induced and stated second-order beliefs.*

Regarding B-subjects' altruism sensitivity, Table 1 shows that, for a given second-order belief, a higher SVO angle increases significantly the likelihood of choosing *Right* in the Reversed-Investment game and not in the Investment or the Exploitation games, as predicted by H.B. 2. In contrast, in the Donation game no effect reaches a standard significance level. It should be noted, however, that Spearman correlations between the SVO angle and the number of *Right* choices for the four induced second-order beliefs are significant and positive both in the Reversed-Investment and Donation games (respectively, $\rho_S = 0.42$, $p = 0.000$; $\rho_S = 0.30$, $p = 0.003$).

Regarding B-subjects' guilt sensitivity, Table 1 reveals a significant impact on the likelihood of choosing *Right* only in the Reversed-Investment and the Donation games, and in the latter only conditional on controlling for personality and socio-demographic characteristics.[20] With this, we conclude that H.B. 2 is broadly supported for altruism sensitivity and only partially for guilt sensitivity, as stated in R.B. 2.

R.B. 2. *[Choice-type correlation] The likelihood of B-subjects choosing* Right *increases in: (i) their altruism sensitivity in the Reversed-Investment game and, to some extent, in the Donation game; (ii) their guilt sensitivity in the Reversed-Investment and Donation games. This holds independently of the treatment.*

H.B. 3 states that guilt aversion is game-independent. To test this, we compare the proportion of B-subjects switching from *Left* to *Right*, that is, exhibiting guilt aversion, across

[19] These results replicate when we regress B-subjects' choices separately on induced second-order beliefs or stated first-order beliefs (see Table D.4 in Section D.3).

[20] The absence of significance in the other two games may not be surprising given that, differently from the measure of altruism sensitivity, our measure of guilt sensitivity was not incentivized (see Bellemare et al., 2019, on the difficulty of finding empirical relationships between the concept of guilt aversion in economics and its characterization in psychological questionnaires).

games. Pooling the treatments, we cannot reject the null hypothesis that the proportion of guilt-averse B-subjects is the same across games (Cochran Q test; $p = 0.509$). Consistently, pairwise comparisons of games reveal no significant difference in the proportion of guilt-averse B-subjects (McNemar tests; lowest p-value, $p = 0.210$).[21] This also holds for each treatment separately (lowest p-value within treatment A, $p = 0.109$; lowest p-value within treatment C, $p = 0.453$). H.B. 3 is thus essentially supported, as stated in R.B.R.B. 3.

R.B. 3. *[Within-subject game-independent guilt] The proportion of guilt-averse B-subjects is not significantly different across the four games, regardless of whether treatments are pooled together or not.*

H.B. 4 states that the proportion of guilt-averse B-subjects in each game does not differ across treatments. To test this, we compare the proportion of B-subjects switching from *Left* to *Right* across treatments. Within a game, we find that the treatment has no significant impact on being guilt-averse in the Investment, the Reversed-Investment and the Exploitation games (Fisher exact tests; smallest $p = 0.810$ for the Investment game).[22] In the Donation game, the proportion of guilt-averse B-subjects is higher in treatment A than in treatment C but the difference does not reach a standard level of significance (Fisher exact test; 36.73% *vs.* 21.28%; $p = 0.118$) (this is consistent with Attanasi et al. (2019b). This analysis confirms the importance of our extension of Battigalli and Dufwenberg (2007) to allow for the emergence of guilt toward a player whose payoff does not depend on B's strategy. This finding is not in line with Bellemare et al. (2017) who detected more guilt among trustees than among dictators: in our three-player games, this should have translated into a higher guilt sensitivity of B-subjects toward a player signaling her intentions through her previous move (A, similar to a trustor in a trust game) than toward the passive C (similar to a recipient in a dictator game). Overall, this analysis supports H.B. 4, as stated in R.B. 4.

R.B. 4. *[Between-subject treatment-independent guilt] The proportion of guilt-averse B-subjects is not significantly different across treatments within each game.*

[21] Given our sample size, the odds ratio and a fixed error probability ($\alpha = 0.05$), we ran a post-hoc power analysis using G*Power (Faul et al., 2009). The highest power is achieved when comparing Investment vs. Donation: 21% ($\beta = 79\%$). By looking at the achieved power as a function of the sample size, we would need 563 B-subjects to obtain a power of 95%. The lowest power is achieved when comparing Reversed-Investment vs. Exploitation: 4% ($\beta = 96\%$). We would need 20 234 B-subjects to obtain a power of 95%.

[22] Given our sample size, the odds ratio and a fixed error probability ($\alpha = 0.05$), a post-hoc power analysis shows that the highest power is achieved when comparing treatments in the Investment game, but only at the 5% level ($\beta = 95\%$). By looking at the achieved power as a function of the sample size, we would need 10198 B-subjects to obtain a power of 95%. The lowest power is achieved when comparing treatments in the Exploitation game: 4% ($\beta = 96\%$). We would need 96642 B-subjects to obtain a power of 95%.

We now turn to H.B. 5, which predicts a higher proportion of selfish B-subjects, that is, those who always chose *Left*, in the Investment and Exploitation games. Pooling the treatments, we indeed find that this proportion is higher in the Investment game than in the Reversed-Investment and Donation games (McNemar tests; $p = 0.012$ and $p = 0.007$, respectively). This proportion is also significantly higher in the Exploitation game than in the Reversed-Investment and Donation games ($p = 0.029$ and $p = 0.015$, respectively). These results also hold if we consider treatment A separately (highest $p = 0.057$ for Investment *vs.* Donation game) but not in treatment C (lowest $p = 0.125$ for Investment *vs.* Donation game). We conclude that H.B. 5 is mostly supported, as summarized in R.B. 5.

R.B. 5. *[Game-dependent altruism] B-subjects' probability of being selfish is significantly higher in the Investment and Exploitation games than in the Reversed Investment and Donation games, both in treatment A and when treatments are pooled.*

Finally, following Bellemare et al. (2011) and Attanasi et al. (2019b), we define a structural econometric model to estimate B-subjects's average guilt sensitivity, γ_{Bjk}, toward player j's beliefs about player k's payoff, with $j, k \in \{A, C\}$ in each of the eight game-treatment combinations. Each B-subject chooses between *Right* and *Left* if *In* for each of the four possible first-order beliefs of j about *Right* if *In* (α_{jB}), in order to maximize his utility as defined by Eq. (8). In this Random Utility Model, λ is the noise parameter that we estimate and U_B essentially follows Eq. (2), for $\alpha_{jB} \in \{0, 1/3, 2/3, 1\}$ and $j \in \{A, C\}$:[23]

$$V_B(\gamma_{Bjk}, \lambda, z|In) \quad = \quad U_B(\gamma_{Bjk}, z|In) + \lambda \cdot \epsilon_B(z|In) \tag{8}$$

A conditional Logit model is used to estimate γ_{Bjk}, the sensitivity corresponding to B's guilt, $\max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\}$ in Eq. (4), while fixing to 1 the "sensitivity" corresponding to B's payoff $\pi_B(z|In)$. Table 2 reports the structural estimates of the mean guilt sensitivity in each game-treatment combination, considering only the B-subjects whose behavior is consistent with our model predictions (i.e., choosing always *Left* or always *Right* regardless of the four α_{jB}, or switching from *Left* to *Right* as α_{jB} increases – see Fig. 6). On average, they represent 87.54% of the B-subjects.

[23] Recall that, differently from Bellemare et al. (2011), in our model B can also be altruistic toward C, with altruism sensitivity measured through the parameter ϕ_{BC} (see Eq. (3)). However, the second component of B's feeling of altruism F_{BC}, that is, C's material payoff π_C, is collinear with B's material payoff π_B, by design. Therefore, we cannot estimate the three coefficients (ϕ_{BC}, γ_{Bjk}, and the coefficient corresponding to π_B) in our utility function while estimating the noise parameter of our random utility model in Eq. (8). Thus, we renounce to estimate ϕ_{BC} in the two games in which it is assumed to be non-null, that is, the Reversed-Investment and Donation games. In the remaining games, this is not an issue because, by design, $\phi_{BC} = 0$ (lexicographic altruism).

Table 2: Structural Estimates of Guilt sensitivity for B-Subjects Disclosing Behavior Consistent with the Model

Game	Treatment A				Treatment C				All
	Inv.	Rev-Inv.	Don.	Exp.	Inv.	Rev-Inv.	Don.	Exp.	
γ_{Bjk}	0.43*** (0.03)	0.45*** (0.06)	0.50*** (0.04)	0.39*** (0.04)	0.34*** (0.05)	0.39*** (0.04)	0.36*** (0.05)	0.36*** (0.05)	0.39*** (0.01)
N Obs.	328	328	344	352	304	344	344	344	2688

When pooling all games and treatments, Table 2 shows that, on average, the B-subjects are willing to pay 0.39 ECU to avoid disappointing their co-player's expectations by 1 ECU. Confidence intervals of the estimated γ_{Bjk} under the eight different specifications always overlap with the confidence interval of our benchmark (All). There is no combination of game and treatment in which the desire to avoid disappointing a co-player is higher or lower than the average. More generally, when comparing the estimated γ_{Bjk} by game and treatment, the confidence intervals always overlap. Therefore, these structural estimates essentially confirm R.B. 3 and R.B. 4.

6 Conclusion

Using four three-player Quasi-Trust mini-games, the current study identified different types of players' vulnerability as potential factors influencing a second-mover's guilt toward the other two players. Its main contribution is to evidence, both empirically and theoretically, the independence of guilt aversion from the vulnerability of the decision-maker's co-players. Theoretically, we proposed a model in which guilt aversion depends neither on the game, nor on the treatment, but rather depends on the role played by the decision maker. We contend that guilt is activated even when the beliefs of the disappointed player do not concern her material payoff but the payoff of a third player; this is a crucial assumption of our model, as in Attanasi et al. (2019b).[24]

Empirically, we found that neither the proportion of guilt-averse second movers, nor the intensity of their guilt aversion differed significantly across the four games (i.e., the four combinations of co-players' vulnerability) and across the two treatments (i.e., guilt toward an active vs. a passive co-player). In particular, second movers exhibited a guilt-averse behavior

[24] Attanasi et al. (2019b) were the first to show that "guilt towards another player can be triggered even when decisions have no direct consequences for that player" (Dufwenberg and Patel (2019, p. 3))

even toward the beliefs of players who were not vulnerable at all. Moreover, the second mover's guilt aversion was triggered even though the first mover's intention was mediated by a passive player. This suggests that observing the intentions of co-players is not a necessary condition to trigger guilt. Guilt aversion was similar toward the expectations of players who were responsible for their choices and toward those of players bearing no responsibility, toward the expectations of players who donated money to a partner and toward those of players who exploited their partner. This study not only reveals the relevance of guilt aversion in games where it had never been tested before. It also supports the notion of guilt as being mainly role-dependent in two-stage games with asymmetric roles.

The insensitivity of the second-mover's guilt aversion to manipulations of the co-player's vulnerability and intentions could, however, also be interpreted as a sign of the subjects' confusion. Yet, the first movers' behavior pleads against this interpretation. Their behavior was game-dependent, in line with our model of lexicographic-altruism in which we assume that the first-mover can feel altruism toward the second mover in the Investment game (in which the passive player is a simple observer) and toward the passive player in the other three games. Since roles were assigned randomly to the subjects, there is no reason to believe that the B-subjects were more confused than the A-subjects.

Alternatively, the way we elicited B-subjects' choices may have reduced the potential impact of the co-players' vulnerability. Indeed, participants were asked to make their choice conditional on four levels of expectations of the other player. This contextualization of choices, traditional when testing belief-based preferences (see, e.g., Khalmetski et al., 2015), has potentially overcome the information provided when introducing the other player's vulnerability. This could be tested by first informing the B-subjects of their co-player's expectations and then, asking them to condition their choices on the different manipulations of their co-players' vulnerability. Finally, in our design, the passive player was always the poorest in the income distribution while the first mover was always the richest. While this feature facilitated comparisons across games for the identification of the role of vulnerability on guilt aversion, it might be interesting to disturb this hierarchy of initial earnings to test how it might affect the intensity of guilt aversion in interaction with vulnerability. These possible extensions are left for future investigations.

References

Amdur, D. and Schmick, E. (2013). Does the direct-response method induce guilt aversion in a trust game? *Economics Bulletin*, 33(1):687–693.

Andreoni, J., Harbaugh, W. T., and Vesterlund, L. (2010). Altruism in experiments. In Durlauf, S. N. and Blume, L. E., editors, *Behavioural and Experimental Economics*, pages 6–13. Palgrave McMillan.

Andrighetto, G., Grieco, D., and Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology*, 6:1413.

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

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

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

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

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

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

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

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

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin*, 115(2):243.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81:145–164.

Bellemare, C., Kröger, S., and Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

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

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

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

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

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

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

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

Buskens, V. and Raub, W. (2013). *Rational choice research on social dilemmas: embeddedness effects on trust*. Russell Sage: New York, NY, USA.

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

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

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–37.

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

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

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

Cox, J. C., Kerschbamer, R., and Neururer, D. (2016). What is trustworthiness and what drives it? *Games and Economic Behavior*, 98:197–218.

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

Dhami, S., Wei, M., and al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*, 167:361–390.

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

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Dufwenberg, M. and Patel, A. (2019). Introduction to special issue on psychological game theory. *Journal of Economic Behavior & Organization*, 167(C)(3):181–184.

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

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

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

Engler, Y., Kerschbamer, R., and Page, L. (2018b). Why did he do that? using counterfactuals to study the effect of intentions in extensive form games. *Experimental Economics*, 21(1):1–26.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

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

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–79.

Ghidoni, R. and Ploner, M. (2020). When do the expectations of others matter? experimental evidence on distributional justice and guilt aversion. *Theory and Decision*, 91:1–46.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19(2):382–393.

Kawagoe, T. and Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102:1–9.

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

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Ockenfels, A. and Werner, P. (2014). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

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

Pelligra, V. (2011). Empathy, guilt-aversion, and patterns of reciprocity. *Journal of Neuroscience, Psychology, and Economics*, 4(3):161.

Pelligra, V., Reggiani, T., and Zizzo, D. J. (2020). Responding to (un) reasonable requests by an authority. *Theory and Decision*, 89(1):1–25.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., and Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, 120(2):142–145.

Yu, H., Shen, B., Yin, Y., Blue, P. R., and Chang, L. J. (2015). Dissociating guilt- and inequity-aversion in cooperation and norm compliance. *Journal of Neuroscience*, 35(24):8973–8975.

A Appendix - Literature

Table A.1 presents a lists of published papers, citing Battigalli and Dufwenberg (2007) with an explicit reference to guilt aversion as a motivation of behavior and including an experiment.[25] It shows that 53.84% of the papers corresponds to trust games while 30.76% corresponds to dictator games. Hence, it also means that only 15.38% of the literature on guilt aversion has investigated other games.

Article	Game
Vanberg (2008)	Dictator
Reuben et al. (2009)	Trust
Ellingsen et al. (2010)	Dictator
Bellemare et al. (2011)	Trust
Chang et al. (2011)	Trust
Charness and Dufwenberg (2011)	Participation
Dufwenberg et al. (2011)	Coordination
Pelligra (2011)	Trust
Attanasi et al. (2013)	Trust
Amdur and Schmick (2013)	Trust
Battigalli et al. (2013)	Sender-Receiver
Beck et al. (2013)	Credence Good
Bracht and Regner (2013)	Trust
Kawagoe and Narita (2014)	Trust
Ockenfels and Werner (2014)	Dictator
Regner and Harth (2014)	Trust
Andrighetto et al. (2015)	Trust
Khalmetski et al. (2015)	Dictator
Yu et al. (2015)	Trust
Attanasi et al. (2016)	Trust
Hauge (2016)	Dictator
Ismayilov and Potters (2016)	Trust
Khalmetski (2016)	Sender-Receiver
Balafoutas and Sutter (2017)	Dictator
Balafoutas and Fornwagner (2017)	Dictator
Bellemare et al. (2017)	Trust & Dictator
Ederer and Stremitzer (2017)	Dictator
Bellemare et al. (2018)	Dictator
Engler et al. (2018a)	Trust
Attanasi et al. (2019a)	Trust
Attanasi et al. (2019b)	Embezzlement
Bellemare et al. (2019)	Trust & Dictator
Dhami et al. (2019)	Public Good
Di Bartolomeo et al. (2019)	Trust
Inderst et al. (2019)	Trust
Morell (2019)	Dictator
Ciccarone et al. (2020)	Dictator
Ghidoni and Ploner (2020)	Lost Wallet
Peeters and Vorsatz (2021)	Prisoner Dilemma

Table A.1: List of published experiments on guilt aversion

[25] This list was compiled based on the authors knowledge of the literature.

B Appendix - Player A's Best-Reply Functions & Hypotheses

In each game, A's best-reply strategy is defined as a function of her first-order belief α_{AB} that B chooses *Right* after *In* and of her sensitivity to altruism toward B, ϕ_{AB}, in the Investment game, and toward C, ϕ_{AC}, in the other three games.

By construction, in each game A's altruism toward player h depends on her belief about B's action after *In*. Let us define the net expected altruism of a player A with altruistic sensitivity ϕ_{Ah} toward player h. It is the difference between her expected altruism when she chooses *In* and her altruism when she chooses *Out*, where α_{AB} is A's first-order belief that B chooses *Right* after *In*:

$$\mathbb{E}_A[F_{Ah}(\phi_{Ah}, z|In)] - F_{Ah}(\phi_{Ah}, Out) = \phi_{Ah} \cdot [\alpha_{AB} \cdot \pi_h(R) + (1 - \alpha_{AB}) \cdot \pi_h(L) - \pi_h(O)] \tag{9}$$

Investment Game. In this game, choosing *In* does not affect C's material payoff while it affects B's expected material payoff. Therefore, by lexicographic altruism, A is altruistic toward B. Relying on A's expected altruism toward B (Eq. (9) with $h = B$), we conclude that A chooses *In* if $20 \cdot \phi_{AB} \cdot (1 - \alpha_{AB}) \geq 0$. Relying on A's expected material payoff, we conclude that A chooses *In* if $\alpha_{AB} \cdot 195 + (1 - \alpha_{AB}) \cdot 155 - 170 \geq 0$, *i.e.*, if $\alpha_{AB} \geq 3/8$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AB} \geq \frac{3 - 8 \cdot \alpha_{AB}}{4(1 - \alpha_{AB})} \tag{10}$$

From Eq. (10) follows that if $\alpha_{AB} \geq 3/8$, then A chooses *In* whatever her altruism sensitivity ϕ_{AB}. For lower first-order beliefs α_{AB}, A chooses *In* only if she is altruistic ($\phi_{AB} > 0$), where the lower the α_{AB}, the higher the altruism sensitivity needed to choose *In*. In particular, a highly-altruistic A ($\phi_{AB} \geq 3/4$) chooses *In* regardless of her first-order belief α_{AB}.

Reversed-Investment Game. In this game, choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (9) with $h = C$), we conclude that A chooses *In* if:

$$\phi_{AC} \cdot (40 \cdot \alpha_{AB} - 15) \geq 0 \tag{11}$$

i.e., if $\alpha_{AB} \geq 3/8$. In this game, choosing *In* does not affect A's material payoff. Therefore, A's best reply function relies only on Eq. (11). With this, if $\alpha_{AB} \geq 3/8$, any altruistic A (*i.e.*, with $\phi_{AC} > 0$) chooses *In* regardless of her sensitivity to altruism, as in the Investment Game. However, differently from the Investment Game, any altruistic A chooses *Out* for $\alpha_{AB} < 3/8$. Eq. (11) also shows that a selfish A ($\phi_{AC} = 0$) is indifferent between *In* and *Out* for each first-order belief α_{AB}: we assume that she chooses *In*. The way we break the tied strategies is motivated by experimental demand effects due to both welfare maximization (see, e.g., Charness and Rabin, 2002) and to the fact that choosing *In* let the game unfold with B's strategy being payoff-relevant for himself and player C. Note that this assumption applies as tie breaking rule in all indifferences in Eqs. (10-13).

Donation Game. In this game choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (9) with $h = C$), we conclude that A chooses *In* if $\phi_{AC} \cdot (40 \cdot \alpha_{AB} + 10) \geq 0$. Relying only on A's expected material payoff, we conclude that A never chooses *In* since $-25 < 0$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AC} \geq \frac{5}{2(1 + 4 \cdot \alpha_{AB})} \tag{12}$$

Thus, a necessary condition for choosing *In* is altruistic enough toward player C, *i.e.*, $\phi_{AC} \geq 1/2$. But this is not sufficient: A's first-order belief of *Right* after *In* must be high enough, with higher ϕ_{AC} compensating for lower α_{AB}. At the limit, for $\alpha_{AB} = 0$, only A's types with $\phi_{AC} \geq 5/2$ choose *In*. Thus, the best-reply behavior of A's (type, belief) pairs in this game is qualitatively similar to the one in the Investment Game, featuring a positive type-belief interaction. However, given A's altruistic type (resp., belief of *Right* after *In*) in both games, a higher belief of *Right* after *In* (resp., altruistic

type) is needed to choose *In* in the Donation Game.

Exploitation Game. In this game choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (9) with $h = C$), we conclude that A chooses *In* if $-25 \cdot \phi_{AC} \geq 0$. Hence, an altruistic A ($\phi_{AC} > 0$) never chooses *In*: differently from the other three games, the altruistic action is *Out*. Relying on A's expected material payoff, we conclude that A chooses *In* if $40 \cdot \alpha_{AB} + 10 > 0$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AC} \leq \frac{2(1 + 4 \cdot \alpha_{AB})}{5} \tag{13}$$

Thus, a necessary condition for choosing *In* is that A is not too altruistic toward player C, *i.e.*, $\phi_{AC} < 2$. But this is not sufficient: A's first-order belief of *Right* after *In* must be high enough, with lower ϕ_{AC} compensating for lower α_{AB}. At the limit, for $\alpha_{AB} = 0$, only A's types with $\phi_{AC} \leq 2/5$ choose *In*. Thus, the best-reply relation between A's type and A's belief in this game is of opposite sign of the one in the Donation Game: given A's belief of *Right* after *In*, a lower altruistic type is needed to choose *In*.

All of the above is summarized in Table B.1. For each of the four Quasi-Trust mini-games, it displays A's best-reply strategies as a function of: (i) her sensitivity to altruism toward B, ϕ_{AB}, in the Investment game, and toward C, ϕ_{AC}, in the other three games; (ii) her first-order belief that B chooses *Right* after *In*, α_{AB}, for the four possible first-order beliefs about B choosing *Right* after *In* that A players can hold in our experiment, *i.e.*, $\alpha_{AB} \in \{0, 1/3, 2/3, 1\}$, as explained in Section 4.

Table B.1: A's predicted behavior depending on her altruism sensitivity ϕ_{Ah} and first-order belief α_{AB}, with altruistic (resp., selfish) strategy in dark grey (resp., light grey) color.

Games	Investment				Rev. Investment				Donation				Exploitation			
α_{AB}	0	1/3	2/3	1	0	1/3	2/3	1	0	1/3	2/3	1	0	1/3	2/3	1
$\phi_{Ah} = 0$	*Out*	*Out*	*In*	*In*	*In*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$(0.00, 0.13)$	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$[0.13, 0.40)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$[0.40, 0.50)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*
$[0.50, 0.68)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*In*	*Out*	*In*	*In*	*In*
$[0.68, 0.75)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*
$[0.75, 0.93)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*
$[0.93, 1.07)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*
$[1.07, 1.47)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*
$[1.47, 2.00)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*In*
$[2.00, 2.50)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*
$[2.50, +\infty)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*In*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*

Given A's sensitivity to altruism, the theoretical predictions in Table B.1 show a positive relationship between the likelihood of the *In* choice and A's first-order belief of *Right* after *In*. This holds regardless of the game. In particular, in the Investment game it holds regardless of A's sensitivity to altruism, in the Reversed-Investment game for all altruistic subjects, in the Donation game only for highly-altruistic subjects, and in the Exploitation game for all subjects but highly-altruistic ones. Considering heterogeneity in A's sensitivity to altruism, we elaborate an hypothesis about A's belief-dependent behavior in each game.

H.A.1. [Choice-belief correlation] The frequency of *In* choices by A-subjects increases in their first-order belief about B-subjects choosing *Right* in each game.

Given A's first-order belief of *Right* after *In*, the theoretical predictions in Table B.1 for the Investment and the Donation games show a positive relationship between the likelihood of the *In* choice and A's sensitivity to altruism. For the Investment game, this only holds under first-order beliefs of a distrustful B ($\alpha \leq 1/3$). For the Donation game, the positive relationship holds for

every positive first-order belief. Conversely, for both the Reversed-Investment and the Exploitation games they show a negative relationship between the likelihood of the In choice and A's sensitivity to altruism. For the Reversed-Investment game, this only holds under first-order beliefs of a distrustful B ($\alpha \leq 1/3$). For the Exploitation game, the negative relationship holds for every first-order belief. Considering heterogeneity in both A's sensitivity to altruism and A's first-order belief of *Right* after *In*, the second hypothesis about A's altruistic type-dependent behavior is as follows:

H.A.2.[Choice-type correlation] The frequency of *In* choices by A-subjects increases in their sensitivity to altruism in both the Investment and Donation games. It decreases in A-subjects' sensitivity to altruism in both the Reversed-Investment and Exploitation games.

In the next two hypotheses, besides heterogeneity in A's sensitivity to altruism, we also make the operational assumption that this sensitivity does not vary too much across games within-subject (*i.e.*, each A-subject has a ϕ_{Ah} in the same interval of Table B.1 for each of the four games). This is required in order to elaborate between-game comparisons in terms of type-dependent behavior, given a low ($\alpha_{AB} \leq 1/3$) or a high ($\alpha_{AB} \geq 2/3$) first-order belief about B choosing *Right*, *i.e.*, given beliefs of a distrustful B or a trustful B, respectively.

For A's choice under beliefs of a distrustful B, we elaborate the next hypothesis by looking at the first two columns ($\alpha_{AB} \in \{0, 1/3\}$) of each of the four game panels of Table B.1. We take the Donation game as the reference (control) since it is the only one where it is possible to identify a subset of altruistic A-types predicted to choose *Out* in that game and *In* in at least another game. In fact, for $\alpha_{AB} = 0$, A chooses *Out* in the Donation game regardless of $\phi_{Ah} < 2.5$, whereas in the other three games there exists a subset of A-types with $\phi_{Ah} < 2.5$, predicted to choose *In*.[26] These types are A-subjects with $\phi_{Ah} \geq 0.75$ in the Investment Game, with $\phi_{Ah} = 0$ in the Reversed-Investment Game, and with $\phi_{Ah} < 0.40$ in the Exploitation Game. Table B.1 shows a similar pattern for $\alpha_{AB} = 1/3$: A-types with $\phi_{Ah} \geq 0.13$, with $\phi_{Ah} = 0$ and with $\phi_{Ah} < 0.93$ are predicted to choose *In* respectively in the Investment, Reversed-Investment and Exploitation Game, but not in the Donation Game. We combine the two sets of predictions for $\alpha_{AB} = 0$ and $\alpha_{AB} = 1/3$ in a unique hypothesis about game-dependent behavior of A-subjects who believe that B would be distrustful after *In*, *i.e.*, that he would more likely choose *Left* ($\alpha_{AB} \leq 1/3$)).

H.A.3. [Choice under beliefs of a distrustful A] For A-subjects thinking that *Left* is the most likely action of B-subjects, the frequency of In choices in the Donation game is lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and slightly-altruistic types; (iii) in the Investment game for slightly-altruistic and highly-altruistic types.

For A's choice under beliefs of a trustful B ($\alpha_{AB} \geq 2/3$), we elaborate the next hypothesis by looking at the last two columns ($\alpha_{AB} \in \{2/3, 1\}$) of each of the four game panels of Table B.1. We take the Investment game as the reference (control) since it is the only one where it is possible to identify a subset of A-types predicted to choose *In* in that game and *Out* in at least another game. In fact, A chooses *In* in the Investment Game regardless of ϕ_{Ah}, whereas in the two of the other games there exists a subset of A-types predicted to choose *Out*. For $\alpha_{AB} = 2/3$, these types are A-subjects with $\phi_{Ah} \geq 1.47$ in the Exploitation Game and with $\phi_{Ah} < 0.68$ in the Donation Game. For $\alpha_{AB} = 1$, these types are A-subjects with $\phi_{Ah} \geq 2.00$ in the Exploitation Game and with $\phi_{Ah} < 0.5\grave{a}$ in the Donation Game. We combine the two sets of predictions for $\alpha_{AB} = 2/3$ and $\alpha_{AB} = 1$ in a unique hypothesis about game-dependent behavior of A-subjects who believe that B would be trustful after *In*, *i.e.*, that he would more likely choose *Right* ($\alpha_{AB} > 1/3$)).

H.A.4. [Choice under beliefs of a trustful A] For As thinking that *Right* is the most likely action of B-subjects, the frequency of In choices in the Investment game is: (i) the same as in the Reversed-Investment game, regardless of the altruistic type; (ii) higher than in the Donation game for selfish and slightly-altruistic types; (iii) higher than in the Exploitation game for highly-altruistic types.

Note that we derived H.A. 1 to H.A. 4 without specifying the treatment (A or C) since, in our experiment, players A are unaware, when they make their choices, of the treatment. Therefore, A's behavior should be treatment-independent.

[26] We are aware from extensive experimental literature eliciting sensitivity to altruism that subjects with $\phi_{Ah} \geq 2.5$ are quite rare in the population. They would be indifferent between keeping 2.5 euros to themselves and giving 1 euro to another player (see, e.g., Andreoni et al., 2010; Bellemare et al., 2008). That is why elaborating H.A.3 for $\phi_{Ah} < 2.5$ is without loss of generality.

C Instructions (Translated from French)

We thank you for participating in this experimental session on decision-making. During this session, you can earn money. The amount of your earnings depends both on your decisions and on the decisions of other participants. At the end of the session, you will receive your earnings in cash in a separate room to preserve the confidentiality of your earnings. The earnings you will receive will include:

- your earnings from today's session
- a €5 fee for showing-up on time to the session.

During the session, some of the transactions are conducted in ECU (Experimental Currency Units).

Please turn off your phone. Communication with the other participants is prohibited during the entire duration of the session. If you have questions during the session, raise your hand or press the red button on the side of your desk and we will come to answer in private.

OVERVIEW OF THE SESSION

In this session, there are two parts. The two parts are completely independent. In each part, one or more of your decisions will be randomly selected by the computer. At the end of the session, you will be informed of your decisions, the decisions of other participants (if they affect your earnings) and their impact on your earnings.

At the end of the session you will be asked to answer a final questionnaire.

FIRST PART: OVERVIEW

In this part, the conversion rate is as follows: 10 ECU = €1.

Roles: At the beginning of the first part, the computer program randomly assigns a role to each participant. You can be either a participant A, a participant B or a participant C. Your role is indicated on your computer screen at the beginning of the first part and you keep the same role throughout this part.

Then, the computer program randomly forms groups of three participants, with one participant of each role in each group. The computer program forms a new group for each situation (which we will describe below), so your group composition changes during the first part. You will never know the identity of the other members of your group and they will never be informed on your identity.

Decisions: Each participant receives an initial endowment. First, Participant A has to make a decision. He can send 25 ECU to Participant B or not. The 25 ECU sent to Participant B come from the endowment of either Participant A or Participant C, depending on the situation.

Then, if Participant B has received 25 ECU, he has to make a decision. He decides how to distribute these 25 ECU between another participant (A or C, depending on the situation) and himself. The ECU that Participant B transfers to another participant (A or C, depending on the situation) are multiplied by two, whereas the ECU that Participant B keeps for himself are not multiplied by two.

Situations: There are four different situations: "North", "West", "East" and "South" (the name of each situation has been given arbitrarily). Decisions are made in each of these four situations.

- In the North situation, Participant A decides whether or not to send 25 ECU from his initial endowment to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant C and himself.
- In the West situation, Participant A decides whether or not to send 25 ECU of his initial endowment to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant A and himself.
- In the East situation, Participant A decides whether or not to send 25 ECU from the initial endowment of Participant C to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant C and himself.
- In the South situation, Participant A decides whether or not to send 25 ECU from the initial endowment of Participant C to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant A and himself.

We will now describe in details the roles, decisions and situations in the first part.

FIRST PART: ROLES, DECISIONS, SITUATIONS

Participant A receives an initial endowment of 170 ECU.

He decides whether or not to send 25 ECU from either his endowment or Participant C's endowment to Participant B.

In the North situation, Participant A has the choice between:

- sending 25 ECU from his initial endowment to Participant B
- sending 0 ECU from his initial endowment to Participant B

In the West situation, Participant A has the choice between:

- sending 25 ECU from his initial endowment to Participant B
- sending 0 ECU from his initial endowment to Participant B

In the East situation, Participant A has the choice between:

- sending 25 ECU from Participant C's initial endowment to Participant B
- sending 0 ECU from Participant C's initial endowment to Participant B

In the South situation, Participant A has the choice between:

- sending 25 ECU from Participant C's initial endowment to Participant B
- sending 0 ECU from Participant C's initial endowment to Participant B

Participant B receives an initial endowment of 100 ECU.

If Participant A has sent 25 ECU to Participant B, Participant B has to make a decision. Then, participant B decides how to distribute these 25 ECU between another participant (A or C, depending on the situation) and himself. The ECU that Participant B transfers to another participant (A or C, depending on the situation) are doubled, whereas the ECU that Participant B keeps for himself are not doubled.

In the North situation, Participant B has the choice between:

- transferring the 25 ECU to the participant C - the participant C receives 50 ECU
- transferring 5 ECU to the participant C - the participant C receives 10 ECU - and keeping 20 ECU for himself - the participant B keeps 20 ECU.

In the West situation, Participant B has the choice between:

- transferring the 25 ECU to Participant A - Participant A receives 50 ECU
- transferring 5 ECU to Participant A - Participant A receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

In the East situation, Participant B has the choice between:

- transferring the 25 ECU to Participant C - Participant C receives 50 ECU
- transferring 5 ECU to Participant C - Participant C receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

In the South situation, Participant B has the choice between:

- transferring the 25 ECU to Participant A - Participant A receives 50 ECU
- transferring 5 ECU to Participant A - Participant A receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

If Participant A has not sent 25 ECU to Participant B, Participant B does not make any decision.

Participant C receives an initial endowment of 30 ECU. Irrespective of the situation, he does not make any decision.

FIRST PART: STAGES

The first part of this session consists of four stages:

- Stage 1: All participants answer some questions.
- Stage 2: Participant A makes his decisions in the four situations.
- Stage 3: Participant B makes his decisions in the four situations.
- Stage 4: Participant A and Participant B answer some questions.

FIRST PART: COMPREHENSION QUESTIONNAIRE

Please complete the comprehension questionnaire that we will distribute to you. If you have any difficulty answering the questionnaire or when you have completed the questionnaire, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

End of the first set of instructions

STAGE 1

In this stage, all participants answer some questions.

If you are a Participant B or a Participant C, you have to answer the following question: "Out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". You have to answer this question for each situation: North, West, East and South.

If you are a Participant A or a Participant C, you have to answer the following question: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". You have to answer this question for each situation: North, West, East and South.

How do the answers to these questions affect your earnings?

At the end of the session, for each role, one of the questions you answered during this stage will be randomly selected by the computer program. If your answer to this question is correct, you earn €1.

Example: Suppose you are Participant C and the randomly selected question is: "In the West situation, out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". The computer program randomly select 3 Participants A among the Participants A in this session. If, in the West situation, "x" Participant(s) A among the 3 Participants A randomly selected has/have decided to send 25 ECU to Participant B, then, your answer is correct if you answered "x".

STAGE 2

In this stage, Participant A makes his decisions.

If you are Participant B or Participant C, you do not make any decision in this stage.

If you are Participant A, you decide whether or not to send 25 ECU to Participant B. You have to make this decision in each situation: North, West, East and South.

Which decision of Participant A determines the earnings of the group members?

At the end of the session, the computer program randomly selects the situation North, West, East or South. The decision made in the randomly selected situation determines the earnings of the group members. At the end of the session, all group members are informed of Participant A's randomly selected decision.

I you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

End of the second set of instructions

STAGE 3

In this stage, Participant B makes his decisions.

If you are Participant A or Participant C, you do not make any decision in this stage.

If you are Participant B, you decide how to distribute the 25 ECU you received between another participant (A or C) and yourself. You have to make this decision in each situation: North, West, East and South. Furthermore, in each situation, you have to make that decision for each possible prediction of Participant *A/C*.[27] To better understand, look at the screen example below. There are two pieces of information that appear in bold fonts on the screen: information on the situation and information on the prediction of Participant *A/C*.

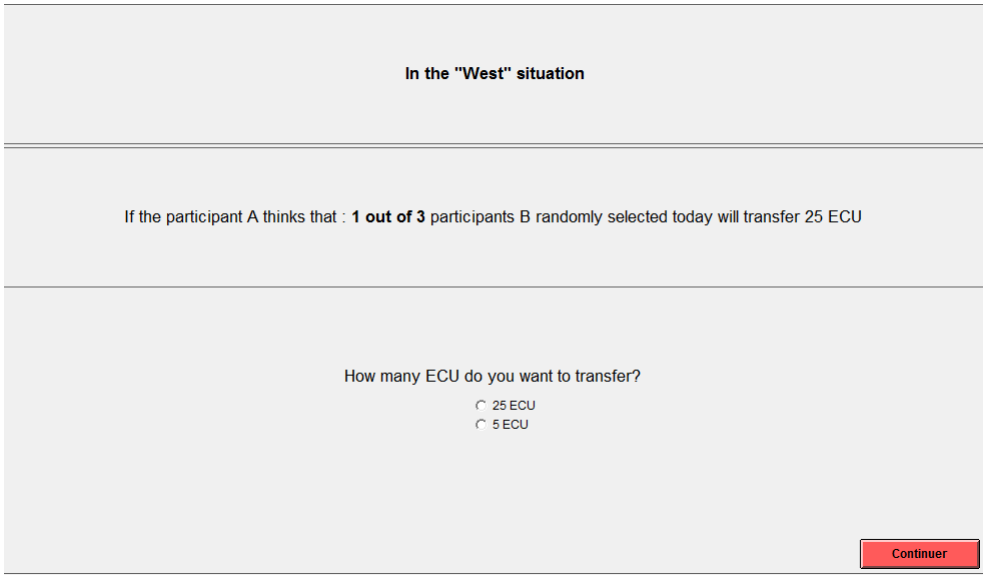

Figure C.1: Screenshot in Treatment A

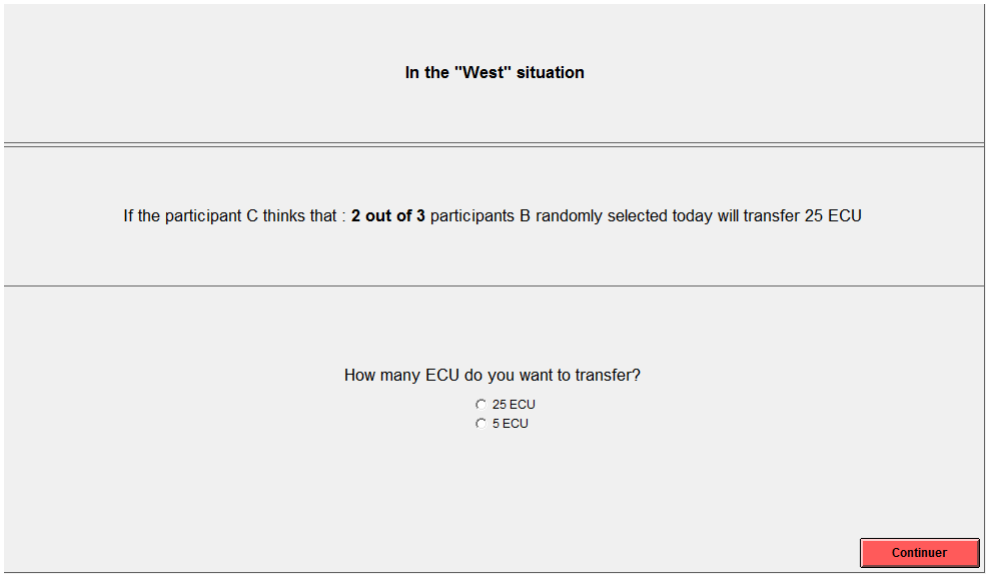

Figure C.2: Screenshot in Treatment C

Information on the situation: You decide how to distribute the 25 ECU in each situation.

Example: In the screen above, you make your decision in the West situation.

[27] Text between *... / ...* represents the two versions of the instructions. The first version corresponds to Treatment A and the second version corresponds to Treatment C.

Information on the prediction of Participant *A/C*: Remember that in stage 1, Participant *A/C* answered the following question for each situation: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". There were four possible predictions: 0, 1, 2 or 3. You decide how to distribute the 25 ECU for each possible prediction of Participant *A/C*.

Example: In the screen above, you make your decision in the West situation, when Participant *A/C* in your group thinks that 2 out of 3 Participants B randomly selected today will transfer 25 ECU to another participant.

To summarize: You must therefore make 16 decisions:

- Four decisions corresponding to the four possible predictions of Participant *A/C* in the North situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the West situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the East situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the South situation

However, only one of these decisions is susceptible to determine the earnings of the group members.

Which decision determines the earnings of the group members?

If Participant A has decided to send 0 EMU to Participant B, no decision of Participant B counts to determine the earnings of the group members.

If Participant A has decided to send 25 EMU to Participant B, a decision of Participant B determines the earnings of the group members. At the end of the session, the computer program randomly selects the situation North, West, East or South. Of the four decisions made by Participant B in the selected situation, the computer program then selects the decision that corresponds to the prediction that Participant *A/C* actually made in stage 1. This decision determines the earnings of the group members.

At the end of the session, all group members are informed of Participant B's randomly selected decision (if any).

Example: Suppose that the computer program randomly selects the West situation. Suppose that, to the question "In the West situation, out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these B Participants will transfer the 25 ECU to another participant?", Participant *A/C* answered "2". Then, the computer program selects the decision that Participant B made when his screen displayed "West situation" and "Participant *A/C* thinks that 2 out of 3 Participants B randomly selected today will transfer 25 ECU to another participant " (see the example screen above). This decision determines the earnings of the group members.

If you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

End of the third set of instructions

STAGE 4

In this stage, Participant A and Participant B answer some questions.

If you are Participant C, you do not make any decisions in this stage. *If you are Participant A*, remember that, in stage 1, Participant B and Participant C answered the following question: "Out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". They answered this question in each situation: North, West, East and South. You have to guess the answers of Participant B and of Participant C in your group.

If you are a Participant B, remember that, in stage 1, Participant A and Participant C answered the following question: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". They answered this question in each situation: North, West, East and South. You have to guess the answers of Participant A and of Participant C in your group.

How do the answers to these questions affect your earnings?

At the end of the session, for each role, one of the questions you answered during this stage will be randomly selected by the computer program. If your answer to this question is correct, you earn €1.

Example: Suppose you are Participant A and the randomly selected question is: "According to Participant C in your group, in the situation West, among 3 Participants A randomly selected in today's session, how

many of these Participants A will send 25 ECU to Participant B?". If, in stage 1, Participant C in your group answered that according to him, in the situation West, "x" Participant(s) A among the 3 Participants A randomly decided to send 25 EMU to Participant B, then, your answer is correct if you answered "x".

If you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

End of the fourth set of instructions

SECOND PART

In this part, the conversion rate is as follows: 10 ECU = €0.1.

There are fifteen periods. In each period, you have to choose the ECU allocation you prefer among nine allocations of ECU that will be proposed to you. An ECU allocation defines how many ECU you receive and how many ECU another participant X, randomly selected, receives.

Your earnings will be determined by one of your choices and by one of the choices of another participant Y, randomly selected. At the end of the session, a period will be randomly selected by the computer program, and the allocations chosen in this period determine your earnings:

- The allocation you have chosen during this period will be implemented for you and for another participant X, randomly selected.
- The allocation that another randomly selected participant Y has chosen during this period will be implemented for you and for him.

Your earnings in the second part are therefore the sum of your payoffs in these two selected allocations.

END OF THE SESSION

At the end of the session, you will be informed of the decisions that will have been selected at random to determine your payoffs (your decisions and those of other participants, if they affect your earnings) and of your final earnings.

Then, you will have to complete a final questionnaire.

At the end of the session, please remain seated and quiet until an experimentalist invites you to proceed to the payment room. Take your computer tag and your payment receipt with you. Leave the instructions on your desk.

If you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

End of the last set of distributed instructions

D Additional Results

D.1 Summary Statistics on Participants, by Treatment

Table D.1: Summary Statistics on Participants, by Treatment

	Treatment A	Treatment C	Treatment Difference
% Women	61.22%	54.61%	No[2]
Mean age	21.90	22.42	No[1]
% Students	94.56%	93.62%	No[2]
% Business major	50.34%	54.61%	No[2]
Mean nb. of past participations	2.07	2.29	No[1]
Mean payoff (€)	17.09	17.03	No[1]
Number of sessions	8	9	
Number of subjects	147	141	

Notes: [1]Mann-Whitney rank-sum test; [2]Fisher exact test; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D.2 Detailed Analysis of A-Subjects Behavior

Table D.2 displays descriptive statistics on the proportion of A-subjects who choose *In*, given their first-order belief (α_{AB}) on the likelihood of *Right* choices by B-subjects.[28]

Table D.2: Proportion of *In* Choices Across Games, by First-Order Belief

% of *In* choices	Inv.	Rev-Inv.	Don.	Exp.
If $\alpha_{AB} = 0$	34.61%	69.76%	4.25%	80.95%
	(52)	(43)	(47)	(63)
If $\alpha_{AB} = 1/3$	48.00%	72.41%	17.85%	60.00%
	(25)	(29)	(28)	(20)
If $\alpha_{AB} = 2/3$	81.00 %	73.33%	64.70%	33.33%
	(16)	(15)	(17)	(6)
If $\alpha_{AB} = 1$	66.66%	66.66%	50.00%	100.00%
	(3)	(9)	(4)	(7)
All (96)	46.87%	70.83%	20.83%	75.00%

Notes: The sample size is in parentheses. α_{AB} is A's first-order belief that B chooses *Right* after *In*.

To estimate the impact of A's first-order beliefs and altruism sensitivity more formally, we report in Table D.3 the results from Logit regressions on the probability that A-subjects choose *In* in each of the four games. For each game, there are two specifications. In the first specification, we regress A-subjects' *In* choice on α_{AB}, that is, their first-order belief on the likelihood of *Right* choices (to test H.A. 1), and on their SVO angle (to test H.A. 2). In line with our assumption of lexicographic altruistic preferences for A-subjects, we consider the SVO angle as a proxy of their altruism sensitivity.

[28]Note that across our 96 A-subjects, the majority (ranging from 45% to 65%) thought that none out of three possible B-subjects would choose *Right* after *In*, and only a small fraction (ranging from 3% to 9%) thought that the three out of three would choose *Right* after *In*.

We control for the treatment and the order of the game. In the second specification, we add personality (self-reported risk aversion and patience) and socio-demographic controls (age, gender, frequency of participation in past experiments, and majoring in business).

Regarding the influence of first-order beliefs (H.A. 1), Table D.3 reports the average marginal effects estimated by random-effects Logit models. It shows that in the Investment and Donation games the more A-subjects believe that *Right* is likely to be chosen by the B-subjects, the more they choose *In*. These marginal effects are highly significant, regardless of the specification. In contrast, in the Reversed-Investment and Exploitation games, first-order beliefs do not significantly influence the frequency of *In* choices. This might be due to the majority of A-subjects being selfish, as usually found for trustees in comparable trust games (see, e.g., Attanasi et al., 2013, 2019a). Recall that in both the Reversed-Investment and Exploitation games, the predicted behavior of selfish A-subjects is *In*, regardless of their α_{AB} (Table B.1). This belief-independent behavior by the bulk of selfish A-subjects could prevent us from detecting the predicted positive correlation for non-selfish ones.

This analysis concludes that H.A. 1 is only supported for the Investment and Donation games. R.A. 1 states that as predicted, the frequency of *In* choices by A-subjects increases in their first-order belief about B-subjects choosing *Right* in the Investment and Donation games, but not in the Reversed-Investment and Exploitation games.

Table D.3: Likelihood of A-Subjects Choosing *In*, by Games

	Investment		Rev-Investment		Donation		Exploitation	
FOB: α_{AB}	0.491***	0.397**	-0.062	-0.125	0.422***	0.463***	-0.103	-0.172
	(0.147)	(0.156)	(0.143)	(0.141)	(0.084)	(0.103)	(0.138)	(0.138)
SVO angle	0.011***	0.009***	0.006	0.005	0.007***	0.007***	-0.004	-0.004
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Treatment A	0.160*	0.197**	0.095	0.069	-0.026	-0.003	-0.050	-0.020
	(0.089)	(0.090)	(0.094)	(0.093)	(0.071)	(0.070)	(0.086)	(0.086)
Order	0.011	0.036	-0.049	-0.042	0.006	0.024	0.085*	0.078
	(0.052)	(0.050)	(0.038)	(0.038)	(0.028)	(0.027)	(0.045)	(0.048)
Personality	No	Yes	No	Yes	No	Yes	No	Yes
Demographics	No	Yes	No	Yes	No	Yes	No	Yes
Observations	96	96	96	96	96	96	96	96
Log-likelihood	-54.361	-48.236	-56.075	-50.634	-33.255	-29.341	-50.963	-46.122
Prob>chi2	0.000	0.000	0.441	0.146	0.000	0.000	0.196	0.107
Pseudo R2	0.181	0.273	0.032	0.126	0.323	0.403	0.056	0.145

Notes: Table D.3 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses. "FOB" for first-order beliefs. "SVO angle" takes value between -7.8 and 45.9. "Order" is the rank order of the game, from 1 to 4. "Personality" controls correspond to the subjects' self-reported risk aversion and patience. "Socio-Demographics" controls include age, gender, frequency of participation in past experiments, and majoring in business. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Regarding the influence of altruism sensitivity on the frequency of *In* choices (H.A. 2), Table D.3 shows that in the Investment and Donation games, the wider is the A-subject's SVO angle, the higher is the likelihood to choose *In*. This effect is highly significant, regardless of the specification. In the Reversed-Investment and Exploitation games, it instead shows no significant effect of A-subject's SVO angle over the likelihood to choose *In*. As for R.A. 1, this absence of significance might be due to the over-representation of selfish A-subjects in our pool, who are predicted to choose *In* regardless of α_{AB}. However, recall that H.A. 2 is meant to show that the motivation behind the *In* choice of A is different across games. Therefore, finding a significant positive correlation between altruism sensitivity and trust in the two games in which this was predicted and no correlation in the two games where a negative relation was predicted provides partial though solid support for H.A. 2.

R.A. 2 states that in the Investment and Donation games, in which a positive choice-type correlation was predicted, there is a significant positive correlation between the frequency of *In* choices by A-subjects and their sensitivity to altruism, whereas in the Reversed-Investment and Exploitation

games, in which a negative choice-type correlation was predicted, the correlation is not significant.

In addition, Table D.3 shows that neither the treatment, nor the order of games have a significant effect on A-subjects' choices, with an exception in the Investment game for the treatment. The two previous results are robust to the inclusion of personality and socio-demographic controls.

In the spirit of the predictions of Table B.1 on the altruism sensitivity ϕ_{Ah}, and of the separation between selfish, lightly-altruistic and highly-altruistic A-subjects on which H.A. 3 and H.A. 4 rely, we split the A-subjects uniformly into these three categories according to their SVO angle. We define as "selfish" the A-subjects with a SVO angle in the interval $(Min, Median - 15\%)$ of the empirical distribution, as "lightly-altruistic" those with a SVO angle in the interval $(Median - 15\%, Median + 15\%)$, and as "highly-altruistic" those with a SVO angle in the interval $(Median + 15\%, Max)$.

To test H.A. 3, we consider the choices of the A-subjects who believe that *Left* is the most likely choice of B-subjects, that is, when A-subjects' $\alpha_{AB} \leq 1/3$. We also rely on the classification of the subjects as selfish, lightly-altruistic and highly-altruistic. Focusing on the selfish A-subjects, we find that the proportion of those who choose *In* is significantly lower in the Donation than in the Reversed-Investment game (2.70% vs. 59.38%; proportion test, $p = 0.000$). Focusing on the selfish and lightly-altruistic A-subjects, we find that this proportion is significantly lower in the Donation than in the Exploitation game (3.77% vs 80.77%; $p = 0.000$). Focusing on lightly-altruistic and highly-altruistic A-subjects, we find that this proportion is significantly lower in the Donation than in the Investment game (15.79% vs 48.89%; $p = 0.001$). These observations also hold if we release the constraints on the A-subjects' SVO angle, that is, if we consider the whole sample of A-subjects (proportion test, $p = 0.000$ for the three comparisons). For the Donation-Investment treatments comparison, this highlights the absence of unpredicted differences in the residual sub-samples of A-subjects, where the predicted choice is *Out* in both games (see Table B.1). For the other two pairwise comparisons, the latter result indirectly confirms within our sample a negligible fraction of highly-altruistic subjects (i.e., those for whom the predicted choice is *In* in the Donation game and *Out* in the other two games).

This analysis supports H.A. 3. R.A. 3 states that, as predicted, for the A-subjects who believe that *Left* is the most likely action of B-subjects, the frequency of *In* choices in the Donation game is significantly lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and lightly-altruistic types; (iii) in the Investment game for lightly and highly-altruistic types.

Finally, to test H.A. 4, we consider the choices of the A-subjects who believe that *Right* is the most likely choice of B-subjects, that is, when A-subjects' $\alpha_{AB} > 2/3$. We find that the proportion of A-subjects who choose *In* in the Investment game is not significantly different than in the Reversed-Investment game, regardless of the altruistic type (proportion test: $p = 1.000$ for selfish, $p = 0.898$ for lightly-altruistic, and $p = 0.177$ for highly-altruistic types). The proportion of selfish and lightly-altruistic A-subjects who choose *In* is higher in the Investment than in the Donation game (60% vs 50%; $p = 0.671$) but not significantly so. Finally, the proportion of highly-altruistic A-subjects who choose *In* is significantly higher in the Investment than in the Exploitation game (100% vs 50%, $p = 0.021$). If we release the constraints on the A-subjects' SVO angle, that is, considering the whole sample of A-subjects, we still find that the proportion of A-subjects choosing *In* in the Investment game is not significantly different than in the Reversed-Investment game (78.94% vs. 70.83%, $p = 0.544$), and higher than in the Donation (61.90%, $p = 0.240$) and the Exploitation games (69.23%, $p = 0.533$).[29] This highlights the absence of unpredicted differences in the residual sub-samples of A-subjects (*i.e.*, with $\phi_{Ah} \in (0.30, 3.33]$), where the predicted choice is *In* in all games (see Table B.1).

This analysis largely supports H.A. 4. R.A. 4 states that, as predicted, for A-subjects who believe that Right is the most likely action of B-subjects, the frequency of In choices in the Investment game is: (i) not significantly different than in the Reversed-Investment game; (ii) higher, although not significantly so, than in the Donation game for lightly-altruistic and selfish types; (iii) significantly higher than in the Exploitation game for highly-altruistic types.

[29] The last difference in proportion is not significant due to the small number of A-subjects with $\alpha_{Ah} \geq 2/3$ in the Exploitation game (only 13/96, see Table D.2).

D.3 Likelihood of B-Subjects Choosing *Right* with induced or stated SOB

Table D.4: Likelihood of B-Subjects Choosing *Right* with stated or induced SOB, by Game

Game	Investment		Rev. Investment		Donation		Exploitation	
Induced SOB: β_{Bj}	2.419***		1.772***		1.685***		3.048***	
	(0.558)		(0.482)		(0.440)		(0.637)	
Stated SOB		3.025***		3.736***		3.222***		2.881***
		(0.786)		(0.883)		(0.873)		(0.737)
SVO Angle	0.046*	0.034*	0.125***	0.083***	0.078***	0.035*	0.058*	0.024
	(0.025)	(0.019)	(0.033)	(0.023)	(0.024)	(0.021)	(0.031)	(0.022)
Guilt Sensitivity	-0.029	0.021	0.204	0.255**	0.038	0.091	0.045	0.059
	(0.110)	(0.087)	(0.128)	(0.102)	(0.099)	(0.086)	(0.146)	(0.098)
Treatment A	0.105	0.114	0.972	1.220**	0.892*	1.324***	-0.398	-0.373
	(0.601)	(0.470)	(0.688)	(0.559)	(0.533)	(0.490)	(0.718)	(0.511)
Order	-0.199	-0.156	0.033	-0.088	0.133	0.052	-0.921**	-0.644**
	(0.324)	(0.258)	(0.266)	(0.212)	(0.205)	(0.176)	(0.391)	(0.264)
Constant	-4.211***	-3.450***	-6.802***	-6.242***	-4.985***	-4.450***	-3.387*	-2.174*
	(1.510)	(1.176)	(1.505)	(1.173)	(1.131)	(0.951)	(1.744)	(1.121)
N Observations	384	384	384	384	384	384	384	384
N subjects	96	96	96	96	96	96	96	96
Log-likelihood	-138.058	-142.061	-165.680	-163.058	-177.317	-178.014	-134.293	-142.339
Wald Chi2	20.89	17.85	24.74	32.61	23.86	24.73	26.50	22.19
Prob>Chi2	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Table 1 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses."Induced SOB β_{Bj}" corresponds to the four β_{Bj} presented to B-subjects when making their choice of *Right* or *Left*. "Stated SOB β_{Bj}" corresponds to the second-order beliefs reported by the B-subjects in the belief elicitation stage. "Reported Guilt" takes values between 0 and 10. "SVO angle" takes values between -7.8 and 38.9. "Order" is the rank order of the game, from 1 to 4. "* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D.4 Within-Individual Analysis of B-Subjects' Consistency of Choices

To explore the within-subject consistency of choices in the four games, we classify the patterns of the B-subjects' decisions into three main categories:

(i) A pattern is said "consistent" when B-subjects always followed the same pattern of choices across the four games (52.08% of B-subjects);

(ii) A pattern is said "nearly consistent" when B-subjects followed the same pattern of choices in three games (20.83% of B-subjects);

(iii) A pattern is said "inconsistent" when B-subjects followed the same pattern of choices in at most two games (27.08% of B-subjects). The details of the choices of inconsistent subjects are available upon request.

The left panel of Figure D.1 displays the distribution of pattern categories for the B-subjects classified as guilt-averse in at least one game. The right panel of Figure D.1 displays the same information for the B-subjects classified as selfish in at least one game.

For both types of preferences, the B-subjects who follow a consistent or nearly consistent pattern of behavior (at least three games) constitute the majority of our observations: 56% of the guilt-averse subjects and 68% of the selfish subjects.

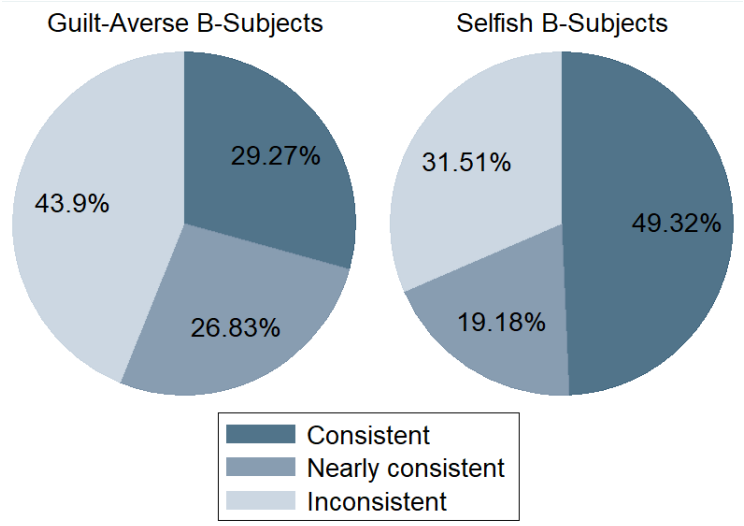

Figure D.1: Distribution of B-Subjects Consistency of Behavior