I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15960

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

Giuseppe Attanasi
Claire Rimbaud
Marie Claire Villeval

FEBRUARY 2023

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15960

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

Giuseppe Attanasi
Sapienza Università di Roma

Claire Rimbaud
University of Innsbruck

Marie Claire Villeval
Univ Lyon, CNRS and IZA

FEBRUARY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

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

We investigate whether a player's guilt aversion is modulated by the co-players' vulnerability. To this goal, we introduce new variations of a three-player Trust game in which we manipulate payoff vulnerability and endowment vulnerability. The former is the traditional vulnerability which arises when a player's material payoff depends on another player's action (e.g., recipient's payoff in a Dictator game). The latter arises when a player's initial endowment is entrusted to another player (e.g., trustor's endowment in a Trust game). Treatments vary whether trustees can condition their decision on the belief of a co-player who is payoff-vulnerable and/or endowment-vulnerable, or not vulnerable at all, and the decision rights of the vulnerable player. We find that trustees' guilt aversion is insensitive to the dimension of the co-player's vulnerability and to the decision rights of the co-player. Guilt is activated even absent vulnerability of the co-player whose beliefs are disappointed. It is triggered by the willingness to respond to the co-player's beliefs on his strategy, regardless of whether this strategy concerns this player or a third player's vulnerability, that is, indirect vulnerability.

JEL Classification: C72, C91, D91

Keywords: guilt aversion, vulnerability, psychological game theory, Dictator game, Trust game, experiment

Corresponding author:
Marie Claire Villeval
Université de Lyon
GATE UMR 5824
93 Chemin des Mouilles
F-69130 Ecully
France
E-mail: villeval@gate.cnrs.fr

* We are grateful to G. Andrighetto, M. Chessa, 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 on their material payoff. Correspondingly, a plethora of psy-game theory-driven experiments have focused on guilt aversion as a driver of prosocial 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 *(i) the co-player's material payoff depends on the actions of the decision maker*. In the Trust game, an additional feature is that *(ii) the use of the co-player's initial endowment, depends on the actions of the decision maker*. In this paper, we refer to these two dimensions as the *vulnerability* of co-players towards the decision maker. Guilt-aversion theory, as originally formulated by Battigalli and Dufwenberg (2007), predicts that the decision maker is concerned only with the expectations of co-players whose payoff can be affected. Hence, it should not matter whether both (i) and (ii) or only (i) characterize the vulnerability of a co-player: the decision maker discloses or not guilt-averse behavior regardless of (ii). Yet, in light of recent experimental evidence discussed below (*e.g.*, Balafoutas and Fornwagner, 2017, Engler et al., 2018b, Attanasi et al., 2019b), whether (and which dimension of) co-players' vulnerability matters is ultimately an empirical question, one that this paper intends to address.

The sensitivity to guilt is likely an intrinsic characteristic of the person, but the context of decision making may leave this emotion dormant or activate it when the individual can harm another person, in particular when this person is vulnerable. The recent literature on guilt aversion has shown that guilt is modulated by a series of factors but it did not explore directly and systematically the role of vulnerability. These studies have focused mostly on the impact of the pre-play communication between players and on the nature of others' expectations. They have shown that pre-play communication increases the trustees' propensity to be guilt-averse, 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

[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 77% of published psy-game experimental studies on guilt aversion (see Table A.1 in Appendix A).

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 prosocial 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).

At the same time, studies have considered vulnerability but not in connection with guilt aversion. In particular, the mediating role of vulnerability has been examined 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' prosocial behavior increases with the vulnerability of the trustor.[3] However, the trustors' vulnerability depends on their decision, and beliefs are not elicited. Hence, these studies are silent on the impact of vulnerability on guilt aversion.

Our work aims to bridge the gap between these two strands of the literature by studying the impact of co-players' vulnerability on guilt. Precisely, we study not only whether the existence but also the dimension of vulnerability matter in the induction of guilt aversion. We distinguish between material payoff and endowment vulnerability. We define a player as **payoff-vulnerable** if *her material payoff depends on the decision maker's choice*. We define a player as **endowment-vulnerable** if *the use of her endowment depends on the decision maker's choice* (*i.e.*, her initial endowment can be entrusted to the decision maker who can then choose how to allocate it). Understanding how payoff vulnerability can trigger guilt is intuitive. Pledging to give to a charity and not actually wiring the money, irrespective of the reason, usually induces the sting of guilt. This payoff vulnerability is a primitive of the traditional formulation of guilt aversion theory, where Battigalli and Dufwenberg (2007) only consider guilt arising from disappointing a co-player's expectations on her material payoff. Yet, psychologists report that "the prototypical cause of guilt would be the infliction of harm, loss, or distress on a relationship partner" (Baumeister et al., 1994, p. 245), that is, a much

[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.

broader definition that includes the guilt from disappointing a co-player's expectations on the use of her endowment that affects a third player's monetary payoff. An obvious, certainly extreme, example is the guilt experienced when not respecting the last wishes of a deceased regarding the allocation of her assets.

Using these definitions, can previous work speak to the issue of vulnerability and guilt aversion? To the best of our knowledge, no final conclusion can be drawn from the existing studies because (i) they all lack a control condition with no vulnerability at all, and (ii) they do not allow for a comparison, in a single frame, of all possible combinations of payoff and endowment vulnerability. On the one hand, by design, one can only observe payoff vulnerability in Dictator games – the dimension of vulnerability emphasized in the traditional formulation of guilt aversion theory. On the other hand, only a joint effect of payoff and endowment vulnerability can be identified in standard Trust games, in which the trustor is both payoff and endowment vulnerable. Only one recent work by Bellemare et al. (2017) contrasted a Dictator and a Trust game in a single study. However, their design does not allow to evaluate the relative impact of each dimension of vulnerability. A first step in this direction has been taken by Attanasi et al. (2019b) who compared guilt aversion toward payoff-vulnerable co-players and endowment-vulnerable co-players in a novel three-player Trust game. In contrast to predictions from the traditional theory of guilt aversion, they detected an impact of endowment vulnerability, and of the same size as that the one of payoff vulnerability. Overall, the existing literature lacks a proper design to compare the relative and joint impact of these two dimensions of co-players' vulnerability on guilt.

With the aim of providing such a comparison, our 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 co-players' vulnerability: no vulnerability, payoff vulnerability, endowment vulnerability, both payoff and endowment vulnerability. This design also offers the possibility to test whether the decision rights of a vulnerable player make a difference in the willingness to avoid disappointing her, by comparing an active player (A) with decision rights, and a passive player (C) with the same dimension of vulnerability but without decision rights. Finally, varying the nature of games allows us to identify the combined effect of co-players' decision rights and dimensions of vulnerability on inducing guilt aversion.

The four Quasi-Trust mini-games are the Investment game, the Exploitation game, the Donation game (similar to Attanasi et al., 2019b), and the Reversed-Investment 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 allocate this money between himself and another player (A or C, depending on the game).[4] Guilt aversion from B to A (resp., C) suggests that the more strongly B believes that A (resp., C) believes that B will allocate this endowment to another player, the more inclined B would be to do so.

The four 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 dimension of vulnerability is activated: payoff and endowment 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 endowment-vulnerable and C payoff-vulnerable, or the reverse). Player B's decisions, 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-subjects. Between-subjects we manipulate whether player B's decisions are elicited conditional on the first-order beliefs of either player A (active) or player C (passive). This 4x2 design allows us to test the (in)dependence of guilt aversion from the co-players' vulnerability and decision rights.

We derive theoretical predictions on belief-dependent behavior of guilt-averse B-subjects in the eight game-treatment combinations. We rely on an extended version of the traditional model of guilt aversion of Battigalli and Dufwenberg (2007) where guilt sensitivity is vulnerability-dependent: B can feel guilty not only if the co-player whose beliefs are disappointed is payoff-vulnerable (as in Battigalli and Dufwenberg, 2007), but also if she is endowment-vulnerable (differently from Battigalli and Dufwenberg, 2007). We still assume that guilt is independent from the co-player's decision rights and that it is null absent any dimension of vulnerability. We complement B's psychological utility by including inequity aversion *à la* Fehr and Schmidt (1999), in order to isolate prosocial behavior not related to guilt aversion but coming from different endowment and payoff distributions across the four games. Although we expect such distribution-dependent preference to play a marginal role in B's behavior, we adopt a conservative approach by allowing B's behavior to depend on other preference specifications that experimental evidence has shown to fit with trustees' and dictators' behavior. Our experimental results confirm a negligible fraction of inequity-averse

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

B-subjects (less than 5%) in the sample, and detect 26% of B-subjects being guilt-averse, twice as many B-subjects with any other non-selfish belief-dependent patterns considered together. This identifies guilt aversion as the main driver of prosocial behavior in the four Quasi-Trust games, thereby highlighting the relevance of our research question.

First, our results generalize those of Attanasi et al. (2019b): the impact of endowment vulnerability on guilt has the same magnitude as that of payoff vulnerability, regardless of the game-treatment combination and decision rights of the disappointed co-player. Second, extending Bellemare et al. (2017), we show that adding a second source of vulnerability of a co-player does not affect guilt sensitivity, again regardless of the underlying strategic interaction and decision rights of the disappointed co-player. Third, and more importantly, guilt aversion is activated even in game-treatment combinations where the disappointed co-player is not vulnerable on any dimension, that is, even when we exclude the dimension of vulnerability (payoff) of the traditional model of Battigalli and Dufwenberg (2007). Guilt sensitivity is in fact independent of the vulnerability and decision rights of the co-player. Indeed, we found no significant difference in the proportion of guilt-averse B-subjects across the four Quasi-Trust mini-games or across treatments, with a relevant fraction of them expressing guilt aversion even toward a co-player who is not vulnerable. The co-players' decision rights do not condition either the players' guilt aversion. To be able to exclude that the insensitivity to the decision rights be driven by B-subjects using player C's (A, respectively) beliefs as informative of player A's (C, respectively) beliefs, we ran an additional treatment. In this treatment, B-subjects could condition their decision on a co-player's beliefs while knowing the true belief of the other co-player. The previous results are robust to this manipulation check.

We interpret such insensitivity of guilt aversion to variations of the co-player's vulnerability and decision rights, while keeping fixed the role and payoff of the potentially guilty player, as further support to guilt triggered by the player's specific role in games with asymmetric roles, as suggested by Attanasi et al. (2016). Once being assigned the role of second mover in a Trust game, a potentially guilt-averse subject discloses consistent belief-dependent behavior regardless of the vulnerability and the decision rights of the co-player he might disappoint. Guilt aversion is triggered by the willingness to respond to a co-player's beliefs on his strategy, regardless of whether this strategy concerns this player or a third player's payoff vulnerability, supporting the notion of indirect vulnerability within the framework of Battigalli and Dufwenberg (2007). To some extent, this reconciles the economic and psychological notions of guilt (Bellemare et al., 2019). An implication of our findings is that guilt aversion is a deeply ingrained human emotion that can be activated in a wider set of circumstances than previously thought.

The remainder of the paper is organized as follows. Section 2 presents our four new games and their rationale. Section 3 introduces our theoretical model and related predictions. Section 4 describes the experimental design. Section 5 and Section 6 present and discuss the results. Section 7 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 Exploitation game (Figure 2), the Donation game (Figure 3), and the Reversed-Investment game (Figure 4). In each game, players A and B are active whereas 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).[5] 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, depending on the game. In the first column (*i.e.*, Figure 1 and Figure 3), the 25 ECU are sent from player A's endowment, whereas in the second column (*i.e.*, Figure 2 and Figure 4), they are sent from player C's endowment. The player from whose endowment the 25 ECU are sent is *endowment-vulnerable*. After *In*, player B decides how to allocate the 25 ECU between himself and player A or C, depending on the game. More precisely, 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.[6] In the first row (*i.e.*, Figure 1 and Figure 2), the ECU can be transferred to player A, whereas in the second row (*i.e.*, Figure 3 and Figure 4), they can be transferred to player C. The player to whom the ECU can be

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

[6] 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. 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.

transferred is *payoff-vulnerable*.

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 endowment-vulnerable and payoff-vulnerable*), but it does not affect C either for her endowment or for her payoff (*i.e., C is non-vulnerable*). The Investment game is a binary version (for the second mover: return 20% *vs.* return 100%) of other Investment games in the literature (see Berg et al., 1995; Ortmann et al., 2000; Buskens and Raub, 2013), with the additional feature of an external observer, C, whose payoff is affected by neither A's, nor B's actions, and who has no decision rights.

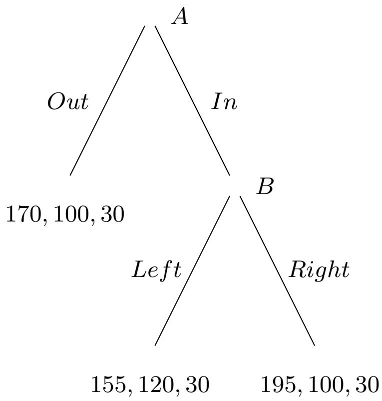

Figure 1: Investment game
ECU: $A \to B \to A$
A vulnerability: endowment & payoff
C vulnerability: none

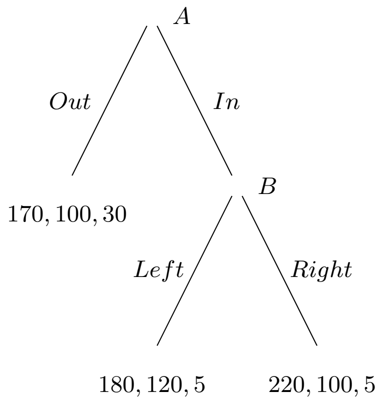

Figure 2: Exploitation game
ECU: $C \to B \to A$
A vulnerability: payoff
C vulnerability: endowment

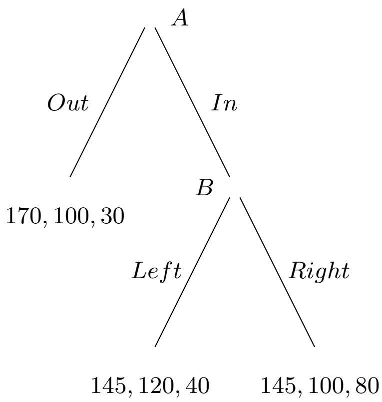

Figure 3: Donation game
ECU: $A \to B \to C$
A vulnerability: endowment
C vulnerability: payoff

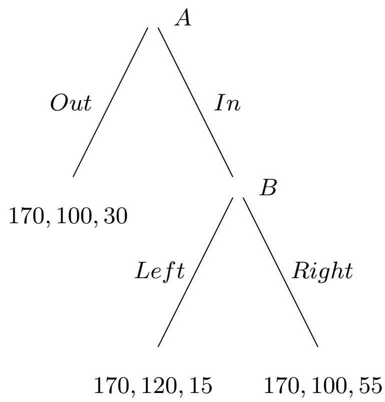

Figure 4: Reversed-Investment game
ECU: $C \to B \to C$
A vulnerability: none
C vulnerability: endowment & payoff

In the **Exploitation 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 A and himself. In this game, B's choice affects both the use of C's initial endowment (*i.e.*, *C is endowment-vulnerable*) and A's material payoff (*i.e.*, *A is payoff-vulnerable*). Thus, this 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.

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 (*i.e.*, *A is endowment-vulnerable*) and C's material payoff (*i.e.*, *C is payoff-vulnerable*). Thus, the Donation game is a modified version of the Investment game where the positive 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 **Reversed-Investment 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 C and himself. In this game, B's choice affects both the use of C's initial endowment and C's material payoff (*i.e.*, *C is endowment-vulnerable and payoff-vulnerable*) but it does not affect A (*i.e.*, *A is non-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.

3 Theoretical Model and Hypotheses

We develop a theoretical model of B's vulnerability-dependent guilt aversion by extending the work of Attanasi et al. (2019b).[7] 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 Quasi-Trust mini-games of Figures 1-4 , that is, respectively, for each terminal history *Out*, *Left* after *In*, and *Right* after *In*.

3.1 Utility Function

Let us introduce B's utility function. Besides B's interest in his own payoff, π_B, we assume that B feels guilty from disappointing j's beliefs on the strategy B will select, with $j \in \{A, C\}$. More precisely, B's disutility from guilt arises from letting down j's beliefs on how

[7] As highlighted in Section 2, the four games have been constructed such that the two dimensions of vulnerability of B's co-players are systematically varied. In our design, understanding the impact of the co-players' different dimensions of vulnerability on the emergence of guilt can only be done through the behavior of player B. In fact, if focusing on player A, she would face a payoff-vulnerable (and endowment non-vulnerable) co-player B regardless of the game, with player C being both payoff-vulnerable and endowment-vulnerable in two out of the four games, which makes an analysis of A's vulnerability-dependent guilt cumbersome.

h's endowment will be used and/or how k's payoff will be affected by B's strategy, with $h, k \in \{A, C\}$ and h not necessarily equal to k:

$$G_B(\gamma_{Bj}, h, k, \alpha_{jB}, z|In) \quad = \quad \gamma_{Bj} \cdot [\mathbb{1}_{\{j\}}(h) + \mathbb{1}_{\{j\}}(k)] \cdot \max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\} \qquad (1)$$

As presented in Eq. (1), B's guilt feeling is the product of three terms:

- $\gamma_{Bj} \geq 0$, B's guilt sensitivity about j's beliefs on B's strategy;
- the sum of the two indicator functions $\mathbb{1}_{\{j\}}(h)$ and $\mathbb{1}_{\{j\}}(k)$, capturing player j's endowment (if $j = h$) and payoff (if $j = k$) vulnerability, respectively.
- 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, $\pi_k(z|In)$.[8]

Define α_{jB} as j's first-order belief that B chooses *Right* after *In*, that is, j's belief that B will act prosocially. Hence, $\mathbb{E}_j[\pi_k(z|In)] = \alpha_{jB} \cdot \pi_k(R) + (1 - \alpha_{jB}) \cdot \pi_k(L)$. If the former is greater than $\pi_k(z|In)$, then B has used h's endowment to increase k's payoff less than j was expecting him to do. In that case, B can feel guilty.

In Eq. (1) we assume that B feels guilty toward j only if j is vulnerable in at least one of the two possible dimensions "endowment" or "payoff", otherwise $\mathbb{1}_{\{j\}}(h) + \mathbb{1}_{\{j\}}(k) = 0$. Furthermore, if B's co-player is both endowment-vulnerable and payoff-vulnerable, then B's guilt feeling is greater ($\mathbb{1}_{\{j\}}(h) + \mathbb{1}_{\{j\}}(k) = 2$) than if the co-player is vulnerable in just one dimension ($\mathbb{1}_{\{j\}}(h) + \mathbb{1}_{\{j\}}(k) = 1$). Note that, absent endowment-vulnerability (*i.e.*, without the term $\mathbb{1}_{\{j\}}(h)$), Eq. (1) reduces to the traditional guilt model of Battigalli and Dufwenberg (2007): player B only feels guilty if the player disappointed in beliefs is payoff-vulnerable ($j = k$). In our model, instead, he can feel guilty also if $j = h \neq k$, that is, if the player whose beliefs are disappointed is endowment-vulnerable although not payoff-vulnerable.

To comment on the implications of guilt and vulnerability in each game, we need to anticipate on the experimental design that we implement (see Section 4). Subjects play the four games of Figures 1-4 under two treatments: in treatment A (respectively, C), player B has to condition his decision only on player A's (respectively, C's) beliefs. Hence, we analyze separately the impact of B's guilt sensitivity toward $j = A$ (in treatment A) from its impact toward $j = C$ (in treatment C), that is, toward a player with (A) or without (C) decision rights in the game. With this, as in Attanasi et al. (2019b), we make the auxiliary assumption that one direction of guilt prevails over the other in each treatment. More precisely, we assume that B can only take into account the beliefs of one co-player, and hence he considers the

[8] This difference also captures how h's endowment is used by player B after In, i.e., how much of it he will transfer to another player, with a prosocial $z = Right$ vs. selfish $z = Left$ end of the Quasi-Trust mini-game.

most salient beliefs (*e.g.*, A's beliefs in treatment A).[9]

Overall, the (traditional) definition of guilt aversion of Battigalli and Dufwenberg (2007) only applies to the Investment and Exploitation games in treatment A (Figures 1-2) and to the Donation and Reversed-Investment games in treatment C (Figures 3-4): only in these four game-treatment combinations, the disappointed player j is payoff-vulnerable. In contrast, our extended definition also applies to the Exploitation game in treatment C (Figure 2, with C endowment-vulnerable) and to Donation game in treatment A (Figure 3, with A endowment-vulnerable). In the remaining two game-treatment combinations – Reversed-Investment game (Figure 4) in treatment A and Investment game (Figure 1) in treatment C – neither of these definitions apply since the disappointed player j is non-vulnerable.

Besides belief-dependent social preferences, we also assume that B may hold belief-independent social preferences, in order to isolate prosocial behavior not related to guilt aversion but coming instead from different endowment and payoff distributions in the mini-games. Thus, we include into B's utility function a distributive-preference component, namely inequity aversion. We model it as aversion to the distance between B's and k's material payoffs, as in Fehr and Schmidt (1999), with B being averse to both disadvantageous inequality ($\max\{0, \pi_k(z|In) - \pi_B(z|In)\}$) and advantageous inequality ($\max\{0, \pi_B(z|In) - \pi_k(z|In)\}$), with $k = A, C$. Since the ranking of payoffs in Figures 1-4 is always $\pi_A(z) > \pi_B(z) > \pi_C(z)$ regardless of players' strategies or games, disadvantageous inequality toward C and advantageous inequality toward A are not possible. Therefore, B's inequity aversion toward A and C reduces to Eq. (2), where, with a slight abuse of notation, ϕ_{BA} represents B's sensitivity to disadvantageous inequality toward A, and $\phi_{BC} \in [0, 1)$, B's sensitivity to advantageous inequality toward C. We assume $\phi_{BA} \geq \phi_{BC}$, as in Fehr and Schmidt (1999).

$$I_B(\phi_{BA}, \phi_{BC}, z|In) = \phi_{BA} \cdot [\pi_A(z|In) - \pi_B(z|In)] + \phi_{BC} \cdot [\pi_B(z|In) - \pi_C(z|In)] \tag{2}$$

Eq. (3) expresses B's overall utility after In, with guilt aversion from Eq. (1) and inequity aversion from Eq. (2), for $j, h, k \in \{A, C\}$:

$$U_B(\gamma_{Bj}, h, k, \phi_{BA}, \phi_{BC}, \alpha_{jB}, z|In) = \pi_B(z|In) - G_B(\gamma_{Bj}, h, k, \alpha_{jB}, z|In) - I_{BC}(\phi_{BA}, \phi_{BC}, z|In) \tag{3}$$

[9] However, B can still form second-order beliefs about the beliefs of more than one co-player, but he will not take them into account for his guilt feeling of Eq. (1). In fact, in an additional study, we empirically test whether knowing the beliefs of A (respectively, C) while conditioning the decision on player C's (respectively A's) beliefs has an impact (see Section 6.2). We essentially find no impact of non-salient beliefs on B's behavior.

3.2 Best-Reply Analysis and Hypotheses

We elaborate our hypotheses on B's behavior relying on best-reply analysis rather than on Bayesian equilibrium.[10] With this, we do not need to introduce A's utility function nor to analyze her best-reply behavior. In fact, B's best-reply strategy according to Eq. (3) does not depend on A's type. It only depends on B's assessment of A's first-order beliefs, which might eventually differ from A's actual beliefs (see the menu method in Section 4).[11]

In each game of Figures 1-4, 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. All transferred amount is doubled. Relying on Eq. (3), we define player B's *Willingness-to-Transfer function* (WT) in Eq. (4) 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 β_{Bj} about the first-order beliefs α_{jB} of the co-player $j \in \{A, C\}$ toward whom he may feel guilty. More precisely, we reason as if player B has a point second-order belief $\beta_{Bj} = \mathbb{E}_B[\alpha_{jB}|In]$ which is conditional on A choosing *In*:[12]

$$\begin{aligned} WT_{j,h,k} &= \mathbb{E}_B[U_B(\gamma_{Bj}, h, k, \phi_{AC}, \phi_{BC}, \alpha_{jB}, R)] - \mathbb{E}_B[U_B(\gamma_{Bj}, h, k, \phi_{AC}, \phi_{BC}, \alpha_{jB}, L)] \\ &= \gamma_{Bjk} \cdot [\mathbb{1}_{\{j\}}(h) + \mathbb{1}_{\{j\}}(k)] \cdot \beta_{Bj} \cdot [\pi_k(R) - \pi_k(L)] - \phi_{BA} \cdot [\pi_A(R) - \pi_A(L)] \\ &\quad + \phi_{BC} \cdot [\pi_C(R) - \pi_C(L)] + (1 + \phi_{BA} - \phi_{BC}) \cdot [\pi_B(R) - \pi_B(L)] \end{aligned} \tag{4}$$

To interpret Eq. (4), with $j, h, k \in \{A, C\}$, consider that the higher player B's willingness to transfer the received endowment, the more player B prefers to choose *Right* rather than *Left*. More precisely, B's best reply is *Right* after *In* if $WT > 0$ and *Left* otherwise. With this, we can find B's best-reply strategy as a function of his second-order belief β_{Bj}, his sensitivity to guilt γ_{Bj} conditional to h's endowment vulnerability and k's payoff vulnerability, and his sensitivity to disadvantageous (resp., advantageous) inequality ϕ_{BA} (resp., ϕ_{BC}).

In the **Investment game** (Figure 1), C is non-vulnerable and A is both endowment-vulnerable and payoff-vulnerable, hence $h = k = A$, and Eq. (4) reduces to:

$$WT_{j,A,A} = 40 \cdot \gamma_{Bj} \cdot [\mathbb{1}_{\{j\}}(A) + \mathbb{1}_{\{j\}}(A)] \cdot \beta_{Bj} - 20 - 60 \cdot \phi_{BA} + 20 \cdot \phi_{BC} \tag{5}$$

[10] 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).

[11] Detailed best-reply analysis and predictions for the behavior of player A can be found in our previous working paper (Attanasi et al., 2022).

[12] 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 using the strategy method, *i.e.*, without B's observing A's decision when making his own.

In treatment $j = A$, $\mathbb{1}_{\{j\}}(A) = 1$, hence guilt is triggered by both dimensions of vulnerability. Hence $WT_{A,A,A}$ in Eq. (5) is strictly positive for all guilt type-belief pairs (γ_{BA}, β_{BA}) such that $\gamma_{BA} \cdot \beta_{BA} > 0.25 + 0.75 \cdot \phi_{BA} - 0.25 \cdot \phi_{BC}$. In treatment $j = C$, guilt is not triggered ($\mathbb{1}_{\{j\}}(A) = 0$), hence $WT_{C,A,A}$ in Eq. (5) is not belief-dependent and γ_{BC} plays no role.

In the **Exploitation game** (Figure 2), C is endowment-vulnerable and A is payoff-vulnerable, hence $h = C$ and $k = A$, and Eq. (4) reduces to:

$$WT_{j,C,A} = 40 \cdot \gamma_{Bj} \cdot [\mathbb{1}_{\{j\}}(C) + \mathbb{1}_{\{j\}}(A)] \cdot \beta_{Bj} - 20 - 60 \cdot \phi_{BA} + 20 \cdot \phi_{BC} \tag{6}$$

Guilt aversion is triggered in both treatments, $j = A$ and $j = C$. In the former, guilt is triggered by A's payoff vulnerability ($\mathbb{1}_{\{j\}}(A) = 1$), in the latter by C's endowment vulnerability ($\mathbb{1}_{\{j\}}(C) = 1$). Hence, independently from $j \in \{A, C\}$, $WT_{j,C,A}$ in Eq. (6) is strictly positive for all guilt type-belief pairs (γ_{Bj}, β_{Bj}) such that $\gamma_{Bj} \cdot \beta_{Bj} > 0.5 + 1.5 \cdot \phi_{BA} - 0.5 \cdot \phi_{BC}$.

In the **Donation game** (Figure 3), as in the Exploitation game, one player is endowment-vulnerable (in this game, $h = A$) and the other is payoff-vulnerable (in this game, $k = C$), hence Eq. (4) reduces to:

$$WT_{j,A,C} = 40 \cdot \gamma_{Bj} \cdot [\mathbb{1}_{\{j\}}(A) + \mathbb{1}_{\{j\}}(C)] \cdot \beta_{Bj} - 20 - 20 \cdot \phi_{BA} + 60 \cdot \phi_{BC} \tag{7}$$

In treatment A, guilt is triggered by A's endowment vulnerability ($\mathbb{1}_{\{j\}}(A) = 1$) and in treatment C by C's payoff vulnerability ($\mathbb{1}_{\{j\}}(C) = 1$), with $WT_{j,A,C} > 0$ in Eq. (7) being strictly positive for all (γ_{Bj}, β_{Bj}) such that $\gamma_{Bj} \cdot \beta_{Bj} > 0.5 + 0.5 \cdot \phi_{BA} - 1.5 \cdot \phi_{BC}$.

In the **Reversed-Investment game** (Figure 4), A is non-vulnerable and C is both endowment-vulnerable and payoff-vulnerable, hence $h = k = C$, and Eq. (4) reduces to:

$$WT_{j,C,C} = 40 \cdot \gamma_{Bj} \cdot [\mathbb{1}_{\{j\}}(C) + \mathbb{1}_{\{j\}}(C)] \cdot \beta_{Bj} - 20 - 20 \cdot \phi_{BA} + 60 \cdot \phi_{BC} \tag{8}$$

In treatment $j = A$, guilt is not triggered ($\mathbb{1}_{\{j\}}(A) = 0$). Hence, $WT_{A,C,C}$ in Eq. (8) is not belief-dependent and the guilt type plays no role. Conversely, in treatment $j = C$, guilt is triggered by both dimensions of vulnerability ($\mathbb{1}_{\{j\}}(C) = 1$). Hence, $WT_{C,C,C}$ in Eq. (8) is strictly positive for all guilt type-belief pairs such that (γ_{BC}, β_{BC}) such that $\gamma_{BC} \cdot \beta_{BC} > 0.25 + 0.25 \cdot \phi_{BA} - 0.75 \cdot \phi_{BC}$.

Therefore, the analysis of $WT_{j,h,k}$ of Eqs. (5)-(8) leads to conclude that, for given sensitivities to inequity aversion, a guilt-averse B is more willing to choose *Right* for higher guilt sensitivity γ_{Bj} and higher second-order beliefs β_{Bj} in all game-treatment combinations apart from: Investment game in treatment C, and Reversed-Investment game in treatment A.

Based on this best-reply analysis, we can derive our hypotheses about B-subjects' belief-dependent behavior. H. 1 and H. 2 consider B-subjects' behavior within each game-treatment combination. H. 3 and H. 4 compare their behavior across game-treatment combinations.

Taken together, H. 1 and H. 2 postulate that guilt is activated in six out of the eight game-treatment combinations of our design. These hypotheses contrast with the predictions from Battigalli and Dufwenberg (2007) and follow-up studies according to which guilt should arise in only four game-treatment combinations, that is, those in which B's strategy is conditioned to the first-order beliefs of a payoff-vulnerable player. In our experimental design, this only occurs in the Investment and Exploitation games (Figures 1-2) under treatment A and the Donation and Reversed-Investment games (Figures 3-4) under treatment C.

Our extended definition of guilt aversion in Eq. (1) instead allows for guilt to be triggered also by conditioning B's strategy to the first-order beliefs of a player who is not payoff-vulnerable. In fact, Attanasi et al. (2019b) have proved experimentally that payoff vulnerability is not a necessary condition for guilt to be activated. They did it in a Donation game comparable to ours of Figure 3, under treatment A. Under this game-treatment combination, however, A is endowment-vulnerable. In our design, we include another game-treatment combination where this should occur, *i.e.*, where B's strategy is conditioned to the beliefs of an endowment-vulnerable player: an Exploitation game (Figure 2) under treatment C.

H. 1. [*Choice-belief correlation*] The frequency of *Right* choices by B-subjects increases in their second-order beliefs about *Right* in all game-treatment combinations apart from the Investment game in treatment C and the Reversed-Investment game in treatment A.

H. 2. [*Choice-guilt type correlation*] Given a positive second-order belief, the frequency of *Right* choices of B-subjects increases with their guilt sensitivity in all game-treatment combinations apart from the Investment game in treatment C and the Reversed-Investment game in treatment A.

After checking that B's belief-dependent and guilt-dependent behavior only occur in those game-treatment combinations where the disappointed player is vulnerable in at least one dimension, we now examine the frequency of such guilt-averse behavior. In this regard, H. 3 and H. 4 posit a fraction of guilt-averse B-subjects in the two game-treatment combinations where the disappointed co-player is vulnerable in two dimensions, which is higher than in the four combinations where she is vulnerable in just one dimension, which is higher than in the two combinations where she is non-vulnerable.

H. 3 and H. 4 explore all the above mentioned comparisons about the detected fraction of guilt-averse B-subjects in a complementary way: H. 3 compares the four games by keeping the treatment fixed; H. 4 compares the two treatments by keeping the game fixed.

H. 3. [*Within-subject game-dependent guilt*] In treatment A, the fraction of guilt-averse B-subjects is significantly greater in the Investment game than in the Exploitation and Donation games, and significantly greater in all the previous games than in the Reversed-Investment game. In treatment C, the opposite prediction holds.

H. 4. [*Between-subject treatment-dependent guilt*] In the Investment and the Reversed-Investment games, the fraction of guilt-averse B-subjects is significantly different across the two treatments A and C. In the Exploitation and Donation games, the fraction of guilt-averse B-subjects is not significantly different across the two treatments A and C.

H. 3 and H. 4 focus on the relevance of guilt-averse B-subjects across the population of participants in that role. For all these subjects, the implicit assumption is that guilt aversion prevails over inequity aversion. We complement the empirical analysis by elaborating an additional hypothesis on the (ir)relevance of inequity aversion across the four games (H. 5).

The behavior of inequity-averse B-subjects is not belief-dependent: within each game-treatment combination, they either always choose *Left* or always choose *Right* for each second-order belief about *Right*. Eqs. (5)-(8) show that B is less willing to choose *Right* for higher ϕ_{BA}, his aversion to disadvantageous inequality toward A. Conversely, B is more willing to choose *Right* for higher ϕ_{BC}, his aversion to advantageous inequality toward C. Therefore, (i) a higher ϕ_{BA} pushes B toward selfish *Left* choice and (ii) a higher ϕ_{BC} pushes B toward prosocial *Right* choice.

Absent guilt aversion, Eqs. (5)-(6) show that the prosocial effect of (ii) prevailing on the selfish effect of (i) is highly implausible in the Investment and Exploitation games. Indeed, for aversion to disadvantageous inequality to prevail over aversion to advantageous one in Eqs. (5)-(6), it must be that $\phi_{BC} > 3 \cdot \phi_{BA}$.[13] This condition is inconsistent with the experimentally validated assumption from Fehr and Schmidt (1999) that $\phi_{BC} \leq \phi_{BA}$, that is, subjects dislike advantageous inequality less than disadvantageous inequality.[14] In contrast, in the Donation and Reversed-Investment games, it is possible that the prosocial effect of (ii) prevails on the selfish effect of (i). From Eqs. (7)-(8), the condition to be satisfied is $3\phi_{BC} > \phi_{BA} + 1$, which is consistent with both $\phi_{BC} \leq \phi_{BA}$ and $\phi_{BC} < 1$.

Therefore, B's inequity-averse behavior is only detectable in the Donation and Reversed-Investment games, where H. 5 states that inequity aversion works as a reinforcement of guilt-averse prosocial behavior. This ultimately leads to a smaller fraction of selfish B-subjects (*i.e.*, always choosing *Left*), with respect to the other two games. In fact, in the Investment and Exploitation games, inequity aversion pushes the selfish choice *Left* which becomes behav-

[13]In the extreme case where aversion to disadvantageous inequality is null ($\phi_{BA} = 0$), the validated assumption from Fehr and Schmidt (1999) that $\phi_{BC} < 1$, *i.e.*, that B cares (at least a little bit) about his own payoff, makes it impossible to satisfy the constraint $\phi_{BC} > 3 \cdot \phi_{BA} + 1$ from Eqs. (5)-(6).

[14]See, *e.g.*, the experimental results of Charness and Rabin (2002), Fehr and Fischbacher (2004), Falk et al. (2008), and a plethora of follow-up tests of the inequity-aversion model of Fehr and Schmidt (1999).

iorally indistinguishable from the behavior of purely self-interested B-subjects. Since inequity aversion is a belief-independent preference, H. 5 should hold regardless of the treatment.

H. 5. [*Game-dependent inequity*] If the fraction of B-subjects who behave selfishly is not constant across games, then it is higher in the Investment and Exploitation games than in the Donation and Reversed-Investment games. This holds independently of the treatment.

Note that observing inequity-averse behavior in the Donation and Reversed-Investment games is empirically implausible, although *a priori* plausible according to Fehr and Schmidt (1999) assumptions. In fact, even by imposing $\phi_{BA} = 0$ in Eqs. (7)-(8), this would only concern B-subjects with $\phi_{BC} \in (1/3, 1)$, and it is well-known that in Dictator and Trust games B-subjects giving more weight to the co-player's than to their own payoff (*i.e.*, with $\phi_{BC} > 1/2$) are rare.[15] In this regard, the doubting tone of H. 5 indirectly highlights the relevance of guilt aversion as the main driver of prosocial behavior in our four mini-games.

4 Experimental Design and Procedures

In our experimental design, each subject went through the four Quasi-Trust mini-games of Figures 1-4 (Investment, Exploitation, Donation and Reversed-Investment) sequentially. 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 sessions, and we varied between-subjects the treatments A and C. The order in which the design is described below follows the timing of the experiment: we present first-order belief elicitation, then, A-subject's decision, B-subject's decision, then, second-order belief elicitation, and finally, elicitation of individual preferences.

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 randomly selected in the session, who would choose *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*.

[15] See, *e.g.*, Bellemare et al. (2017), Bellemare et al. (2018), and Attanasi et al. (2019a).

Similarly, they had to report, for each game, their belief about the number of B-subjects, out of three randomly selected in the session, who would choose *Right* conditional on the 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 made their decisions 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. 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).[16] It allows to rule out potential false-consensus effects without raising the issue of strategic reporting and without using outright deception ("unexpected data use" is judged the least deceitful practice by economic researchers, see Charness et al., 2022). 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).[17]

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 and C-subject in the game. In other words, A had to guess B's and C's first-order beliefs on the

[16] Note that the menu method differs from Ellingsen et al. (2010), where the decision maker is shown the true belief of the co-player before making his decision, whereas in the menu method he is shown all possible beliefs of the co-player, and he makes one decision for each of these beliefs. However, similarly as in Ellingsen et al. (2010) where the co-player is not told that her belief will be shown to the matched decision maker, in the menu method the co-player is not told that the decision maker conditions his decisions on her beliefs.

[17] One may be worried that the menu method creates an experimenter demand effect, but Bellemare et al. (2017) concluded that in the Trust game the "menu" approach yields results that are similar to the "baseline" approach (*i.e.*, when decisions are elicited unconditionally).

frequency of A-subjects choosing *In*. We also elicited, for each game, B-subjects' second-order beliefs 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.[18] Finally, we collected socio-demographic characteristics, including gender, age, major and number of past participations 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.1 shows that the mean individual characteristics are similar across treatments.

The session consisted of two parts. The instructions (see Appendix B) for the first part were distributed before each stage. The first stage described the four games. The experimenter made sure that all subjects had completed correctly a comprehension questionnaire

[18]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.

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, patience 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

Since we are mainly interested in B's behavior and in line with the best-reply analysis of Section 3.2, we focus here on B-subjects' behavior. The results on A-subjects' behavior are available in Appendix C, which essentially show that it differs across games but not across treatments. Except when specified otherwise, the non-parametric tests are two-sided.

Before testing our hypotheses H. 1 – H. 5 formally, B-subjects' behavior can be classified into five patterns of choices through their four induced second-order beliefs, $\beta_{Bj} \in \{0, 1/3, 2/3, 1\}$, in each game:[19]

(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;[20]

(ii) always choosing *Right*, regardless of the induced second-order beliefs, that is, disclosing inequity aversion: 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;

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

[20] 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.

(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 5 displays the distribution of B-subjects across these patterns of choices in all game-treatment combinations.[21] Among the five patterns of behavior identified above, our model is consistent with behaviors described in patterns (i), (ii) and (iii) (selfish, inequity-averse and guilt-averse), which 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 in our games.

We next directly test our hypotheses. Table 1 presents the average marginal effects from panel Logit regressions on the probability to choose *Right*. We regress B-subjects' choices on their induced and stated second-order beliefs as well as their self-reported guilt proneness, γ_{Bj}. We control for the treatment and the order in which the game was played. We also include personality (prosociality, risk aversion and patience) and socio-demographic controls (age, gender, frequency of past participation in experiments, majoring in business).

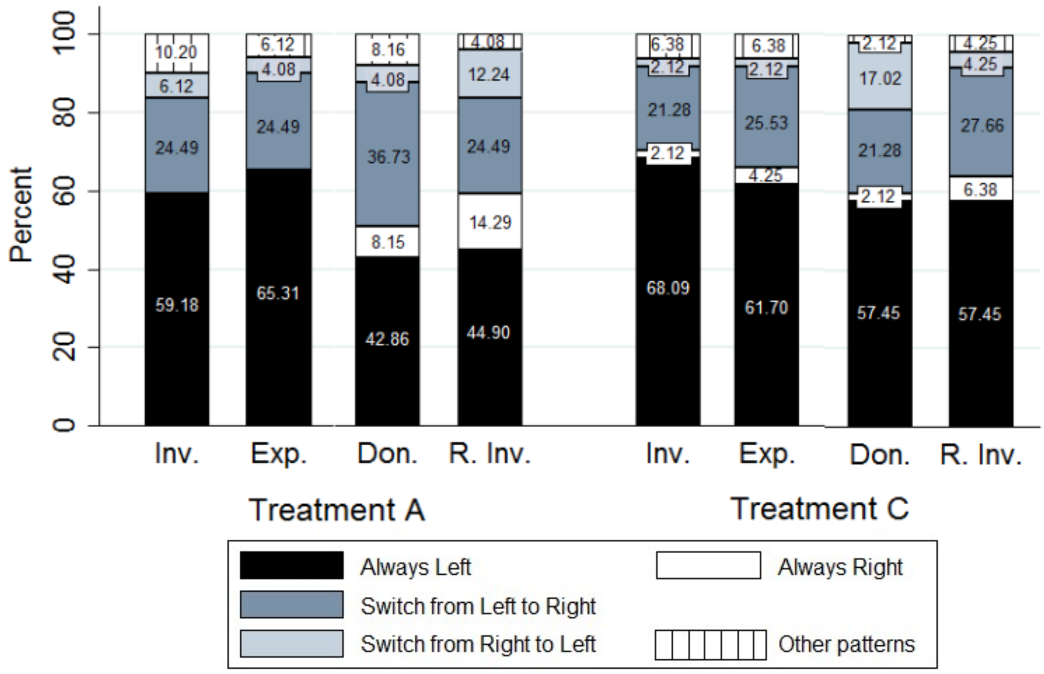

Figure 5: Distribution of B-subjects' patterns of choices across games and treatments

Table 1 shows that, in almost all game-treatment combinations, the higher is the induced second-order belief, the more likely B-subjects are to choose *Right*. The same holds for the

[21]In addition, Figure D.1 in Appendix D.2 analyzes the consistency of B-subjects' patterns of choices across the four games.

stated second-order beliefs. One exception is in the Donation game in treatment C, where this non-significant effect can be explained by the high proportion of B-subjects switching from *Right* to *Left*, that is, in the opposite direction from guilt-averse subjects.[22] We conclude that H. 1 is essentially supported.

R. 1. *[Choice-belief correlation] The frequency of B-subjects' Right choices increases with their second-order beliefs in five out of six game-treatment combinations where it was predicted to do so.*

Table 1: Likelihood of B-subjects choosing *Right*, by game-treatment combination

Game	Investment		Exploitation		Donation		Rev. Investment	
Treatment	A	C	A	C	A	C	A	C
Vulnerability	both	no	pay.	end.	end.	pay.	no	both
Induced SOB	0.204***	0.194***	0.234***	0.203***	0.325***	0.032	0.118*	0.221***
	(0.064)	(0.052)	(0.057)	(0.053)	(0.062)	(0.063)	(0.065)	(0.057)
Stated SOB	0.303***	0.274**	0.245***	0.287***	0.420***	0.427***	0.533***	0.203
	(0.085)	(0.114)	(0.065)	(0.087)	(0.132)	(0.140)	(0.084)	(0.128)
Reported Guilt	-0.008	0.007	-0.009	0.030*	0.028*	0.027*	0.029**	0.026
	(0.010)	(0.014)	(0.010)	(0.016)	(0.016)	(0.016)	(0.014)	(0.019)
Order	-0.070**	0.033	-0.061**	-0.044	0.017	-0.067**	-0.005	0.027
	(0.031)	(0.045)	(0.024)	(0.031)	(0.032)	(0.026)	(0.041)	(0.033)
Personality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N Observations	196	188	196	188	196	188	196	188
N subjects	49	47	49	47	49	47	49	47
Log-likelihood	-68.514	-48.977	-56.255	-55.118	-82.030	-66.538	-82.043	-64.452
Wald Chi2	21.92	14.95	23.61	17.34	24.08	25.32	24.56	17.67

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

Table 1 also reveals a significant impact of self-reported guilt sensitivity on the likelihood of choosing *Right* in four game-treatment combinations.[23] However, whether the impact of guilt sensitivity is significant or not is not related to the vulnerability of co-players. Hence, we conclude that H. 2 is only partially supported.

R. 2. *[Choice-guilt type correlation] The frequency of B-subjects' Right choices increases with their guilt sensitivity in three out six game-treatment combinations where it was predicted to do so.*

[22]These results replicate when we regress B-subjects' choices separately on induced second-order beliefs or stated second-order beliefs (see Table D.2 in Appendix D.3).

[23]The absence of significance in the other combinations may not be surprising given that 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).

We compare the proportion of guilt-averse B-subjects across games. Independently from the treatment, we cannot reject the null hypothesis that this proportion is the same across games (Cochran Q tests; p-value = 0.133 in treatment A, p-value = 0.556 in treatment C). Consistently, pairwise comparisons of games reveal no significant difference in the proportion of guilt-averse B-subjects (McNemar tests; lowest p-value = 0.109 in treatment A, p-value = 0.453 in treatment C).[24] Overall, H. 3 is not supported in our data.

R. 3. *[Within-subject game-dependent guilt] The proportion of guilt-averse B-subjects is not significantly different across the four games, irrespective of the treatment.*

Next, we compare the proportion of guilt-averse B-subjects 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; lowest p-value = 0.118 for the Donation game).[25] Since guilt-averse B-subjects have strictly positive guilt sensitivity, we also tackle this question by comparing their guilt sensitivity measured by the switching second-order belief in each treatment (see Figure D.2 in Appendix D.4). Keeping the game constant, the distributions of switching SOB do not differ significantly across treatments, except for the Donation game (Kruskal-Wallis test, p-value = 0.029). Overall, H. 4 is not supported in our data.

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

We also regressed B-subjects' choice to go *Right* on interaction terms between induced SOB and co-players' vulnerability, as well as between induced SOB and decision rights (*i.e.*, the treatment). We found that subjects mostly did not react differently to the induced SOB of another player depending on this other player's vulnerability (see Table D.3 in Appendix D.5). These results confirm R. 3 and R. 4.

We now turn to H. 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 Donation and Reversed-Investment games (McNemar tests; p-value = 0.007 and p-value = 0.012, resp.).

[24] Given our probability of discordant pairs, the odds ratio and a fixed error probability ($\alpha = 0.05$), we ran an a priori power analysis using G*Power (Faul et al., 2009). Considering the comparison which was the closest to significance (Donation *vs.* Exploitation in Treatment A, and Donation *vs.* Reversed-Investment in Treatment C), the analyses concluded that, to achieve power 80%, 669 participants would be required in Treatment A, and 12,226 participants in Treatment C.

[25] Given the probabilities to be guilt-averse and a fixed error probability ($\alpha = 0.05$), we ran an a priori power analysis using G*Power (Faul et al., 2009). Considering the comparison which was the closest to significance (in the Donation game), the analysis concludes that, to achieve power 80%, 843 participants would be required.

This proportion is also significantly higher in the Exploitation game than in the Donation and Reversed-Investment games (p-value $= 0.015$ and p-value $= 0.029$, resp.). 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-value $= 0.125$ for Investment *vs.* Donation game). We conclude that H. 5 is mostly supported, as summarized in R. 5.

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

6 Discussion of the Results

6.1 A Structural Estimation of a Model of Indirect Vulnerability

We find that B's guilt aversion is insensitive to the dimension of the co-player's vulnerability. Guilt is even activated absent any dimension of vulnerability of the co-player whose beliefs are disappointed. Based on these results, it appears that guilt aversion is triggered by the willingness to respond to the co-player's beliefs on his strategy, regardless of whether this strategy concerns this player's or a third player's vulnerability (*i.e.*, **indirect vulnerability**). To better fit the data, we propose to re-define player B's guilt in Eq. (9). With respect to our model of Eq. (1), B's disutility from guilt in Eq. (9) includes no indicator function for either payoff or endowment vulnerability of the co-player. Therefore, with respect to the traditional guilt model of Battigalli and Dufwenberg (2007), rather than including a further dimension of vulnerability (endowment) of the disappointed co-player, Eq. (9) pulls away even this traditional dimension of vulnerability (payoff). With this, player B can feel guilty toward a co-player $j \in \{A, C\}$ who formed beliefs about his strategy, which affects the material payoff of player $k \in \{A, C\}$, with j not necessarily equal to k:[26]

$$G_B(\gamma_{Bj}, \alpha_{jB}, z|In) \quad = \quad \gamma_{Bj} \cdot \max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\} \tag{9}$$

Based on this last definition of guilt aversion, we define a structural econometric model to estimate B-subjects' average guilt sensitivity (γ_{Bj}) toward player j's beliefs about B's prosocial *Right* choice in each of the eight game-treatment combinations. In line with the menu method of our experimental design, we assume that each B-subject chooses between *Right* and *Left* for each of the four possible first-order beliefs of j about *Right* ($\alpha_{jB} \in \{0, 1/3, 2/3, 1\}$), in order to maximize his utility as defined by Eq. (10). In this Random

[26] Note that in Eq. (9) the material payoff of player k turns out to be just one of the possible ways to measure the difference between player B's expected and actual prosocial behavior: this can be measured through direct ($j = k$) or indirect ($j \neq k$) vulnerability of the disappointed co-player.

Utility Model, we include B's material payoff, his disutility from guilt from Eq. (9), and a noise term:[27]

$$V_B(\gamma_{Bj}, \alpha_{Bj}, \lambda, z|In) \quad = \quad \pi_B(z|In) - G_B(\gamma_{Bj}, \alpha_{jB}, z|In) + \lambda \cdot \epsilon_B(z|In) \tag{10}$$

As in Bellemare et al. (2011), a conditional Logit model is used to estimate guilt sensitivity γ_{Bj} and the noise parameter λ, while fixing to 1 the "sensitivity" corresponding to B's material payoff. Table 2 reports the structural estimates of the mean guilt sensitivity in each game-treatment combination, first considering only B-subjects with the behavioral patterns consistent with Eq. (10) (guilt-averse and selfish B-subjects, on average, 77.25% of the B-subjects)[28] and then all B-subjects.

Table 2: Structural estimates of guilt sensitivity for B-subjects in the eight game-treatment combinations

	Treatment A				Treatment C				All
Game	Inv.	Exp.	Don.	R.-Inv.	Inv.	Exp.	Don.	R.-Inv.	
				Guilt-averse and selfish B-subjects					
γ_{Bj}	0.43***	0.39***	0.50***	0.44***	0.38***	0.38***	0.36***	0.41***	0.41***
	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.01)
N Obs.	328	352	312	272	296	328	336	320	2544
				All B-subjects					
γ_{Bj}	0.31***	0.34***	0.48***	0.25***	0.30***	0.32***	0.06***	0.34***	0.31***
	(0.05)	(0.04)	(0.05)	(0.10)	(0.05)	(0.05)	(0.10)	(0.06)	(0.02)
N Obs.	392	392	392	392	376	376	376	376	3072

Pooling all games and treatments, Table 2 shows that, on average, B-subjects are willing to pay 0.31 ECU to avoid disappointing their co-player's expectations by 1 ECU (0.41 ECU if focusing only on B-subjects consistent with the model). More importantly, regardless of the subject pool, the confidence intervals of the estimated γ_{Bj} almost always overlap with the confidence interval of our benchmark ("All" column of Table 2), that is, the desire to avoid disappointing a co-player is neither higher nor lower than the average in those combinations.[29]

[27] Recall that, differently from Bellemare et al. (2011), in our main model of Eq. (1) player B can also be inequity-averse. However, player A and C's material payoffs (used to measure player B's disadvantageous and advantageous situation, respectively) are collinear with player B's material payoff, by design. Therefore, we cannot estimate the four coefficients (ϕ_{BA}, ϕ_{BC}, γ_{Bj}, and the coefficient corresponding to π_B) in our utility function while estimating the noise parameter of our random utility model in Eq. (10). Thus, given the goal of our econometric exercise, we renounce to estimate the sensitivity to inequity.

[28] Structural estimates considering behavioral patterns consistent with Eq. (3), *i.e.*, including also inequity-averse subjects, are available in Table D.4 of Appendix D.6.

[29] Exceptions are in the Donation game, treatment A, irrespective of the sample, where the estimate is significantly higher than the average; as well as in the Donation game, treatment C, with the whole sample, where the estimate is significantly lower than the average (probably due to the high proportion of B-subjects exhibiting a reverse pattern compared to guilt aversion; see Figure 5).

6.2 Robustness Follow-Up Study

Finally, we ran a follow-up study to rule out potential confounding effects.[30] In our original study, we observed whether B-subjects conditioned their decisions on another player's beliefs. Yet, it is possible that B-subjects use the belief of A in treatment A to infer the true belief of C, and vice versa in treatment C. If that was the case, we could have classified some B-subjects as guilt-averse toward A while they were in fact trying to condition their decision on C's belief (and vice versa).

To test this alternative interpretation of our results, we ran additional experimental sessions where, in treatment A, B-subjects were asked to condition their decision on A's beliefs *while knowing the true belief of C* (and vice versa in treatment C). If the conditionality on A's beliefs that we observed in the original study was only a proxy for γ_{Bj} toward C, we should not observe such conditionality in this new treatment. Since the most novel combinations of our experimental design are those where B-subjects can condition their decision on the beliefs of a player who is not payoff-vulnerable to B's strategy (*i.e.*, the combinations where the traditional theory of guilt aversion predicts no conditionality), we ran within-subjects this additional study only for those game-treatment combinations: (Investment, C), (Exploitation, C), (Donation, A), (Reversed-Investment, A). We invited 150 subjects from the same GATE-Lab subject pool as in the original study and ran seven sessions.

Our main conclusions are robust to this manipulation. Table 3 summarizes the proportion of guilt-averse B-subjects in this additional study and in the corresponding game from the original study. We found no significant differences. A more detailed decomposition of patterns of choices in the follow-up study can be found in Figure D.3 (Appendix D.7).

Table 3: Proportion of guilt-averse B-subjects, by treatment

	Original Study Conditional on ...		Additional Study ... and showing		Difference (test)
Investment	C's possible beliefs	21.28%	A's belief	26.00%	0.585
Exploitation	C's possible beliefs	25.53%	A's belief	24.00%	0.861
Donation	A's possible beliefs	36.73%	C's belief	30.00%	0.477
Rev.-Investment	A's possible beliefs	24.49%	C's belief	28.00%	0.691
N Obs. Treat. (C;A)	(47;49)		(50;50)		

Notes: Last column: p-values of proportion tests: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

[30]We thank an anonymous referee for this suggestion.

7 Conclusion

Our study uses four three-player Quasi-Trust mini-games to identify whether and how different dimensions of players' vulnerability (payoff and endowment vulnerability) influence a second mover's guilt. Payoff vulnerability of a disappointed player is a necessary condition for guilt to be triggered in the traditional model of guilt aversion of Battigalli and Dufwenberg (2007). Attanasi et al. (2019b) were the first to raise doubts on this restriction, by identifying guilt aversion of similar size toward co-players' payoff vulnerability and endowment vulnerability. To compare more systematically the relative and joint impact of these two dimensions of vulnerability on guilt, our design extends the one in Attanasi et al. (2019b) in several directions. First, we test whether their equivalence result between payoff and endowment vulnerability in guilt activation in the Donation game also holds in three other comparable games. Second, we check whether, given the same dimension of vulnerability (*e.g.*, endowment), the same fraction of guilt-averse second movers is found between two comparable games (*e.g.*, the Exploitation game in treatment C and the Donation game in treatment A). Third, we compare game-treatment combinations where guilt can be triggered by both dimensions of vulnerability (payoff and endowment) *vs.* only one dimension (payoff or endowment), implementing a thorough analysis of incremental vulnerability. Our design also extends Bellemare et al. (2017) who tested the impact of incremental vulnerability on guilt by comparing a classical Trust mini-game (with payoff and endowment vulnerability of the trustor) to a classical Dictator mini-game (with only payoff vulnerability of the recipient). Finally, by also proposing two game-treatment combinations where the second mover's strategy is conditioned to the beliefs of a non-vulnerable player, we are able to test whether vulnerability itself, regardless of its dimension, is a necessary condition for guilt activation.

We elaborated a model of vulnerability-dependent guilt in the four Quasi-Trust mini-games under two treatments. This model assumes that the second mover's guilt is activated by either payoff or endowment vulnerability of the co-player and that the combined effect of the two dimensions is positive. Our experimental results contribute to the literature on guilt aversion in the Trust and the Dictator games in three directions.

First, extending Attanasi et al. (2019b), we confirm that the impact of endowment vulnerability is of the same size as the one of payoff vulnerability, regardless of the underlying strategic interaction (game-treatment combination) and decision rights of the disappointed co-player. Second, extending Bellemare et al. (2017) and rejecting one prediction of our model, we confirm that given one dimension of vulnerability of the co-player (*e.g.*, endowment), adding another dimension (*e.g.*, payoff) has no impact on guilt sensitivity, regardless of the underlying strategic interaction and decision rights of the disappointed co-player. Third, more importantly, we detect guilt aversion even in game-treatment combinations where the disappointed co-player is not vulnerable at all. The latter result reconciles the previous two, and so, the previous evidence: the reason why payoff and endowment vulnerability have the same effect on the second mover's guilt aversion (Attanasi et al., 2019b) while their additional combined effect is null (Bellemare et al., 2017) is because his guilt sensitivity is independent of the co-player's vulnerability and decision rights. This can explain why we find a similar proportion of guilt-averse second movers in each of our eight game-treatment combinations.

We confirm this interpretation through a structural estimation of the second mover's average guilt sensitivity by relying on a model that "pulls away" even the dimension of (payoff) vulnerability of the traditional model of Battigalli and Dufwenberg (2007). In this alternative model, guilt aversion is triggered by the willingness to respond to a co-player's beliefs on his strategy, regardless of whether this strategy concerns this player or a third player's vulnerability (*i.e.*, indirect vulnerability). With this, the structural estimates of mean guilt sensitivities are similar in each of the eight game-treatment combinations. These results indicate that our initial model adding new dimensions of vulnerability is less general than this alternative model of indirect vulnerability, the latter including also the two game-treatment combinations previously defined as "no vulnerability." Moreover, an additional robustness study allows us to rule out that our main conclusions rely on potential confounding effects due to the fact that second movers might use the belief of the co-player to which they are asked to condition their decision to infer the true belief of the other co-player.

Overall, on the one side our results highlight the relevance of guilt aversion *à la* Battigalli and Dufwenberg (2007) in games where it had never been tested before. This further shows that guilt aversion is a deeply ingrained human emotion that can be activated in a wider set of circumstances than previously thought (see the review in Battigalli and Dufwenberg, 2022). On the other side, given the insensitivity of guilt aversion to any game and treatment manipulation while keeping fixed the role and payoff of the potentially guilty subject, they indirectly support the notion of guilt as being driven by the player's role in games with asymmetric roles (Attanasi et al., 2016). Once being assigned the role of second mover in a Trust game, a potentially guilt-averse subject discloses consistent belief-dependent behavior regardless of the vulnerability and the decision rights of the co-player he might disappoint. To some extent, this reconciles the economic and psychological notions of guilt (Bellemare et al., 2019): the role played in a strategic interaction is part of the categorization process of the anticipated disappointment of the other that leads an individual to experience the emotion of guilt (*e.g.,* Barrett, 2006).

Some reflections on the limitations of our study open avenues for further extensions. We only focused on vulnerability-dependent guilt of the second mover. However, the first mover in a Trust game may also feel guilty (see, *e.g.*, Attanasi et al., 2016). Exploring whether her guilt sensitivity is also independent of the co-players' vulnerability would generalize our finding, by making it independent from the specific role in the game. In that case, it could be assessed whether indirect vulnerability is a sufficient condition for guilt activation, regardless of the role of the guilty subject. Furthermore, in our design, the passive player was always the poorest in the initial endowment 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.

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.

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.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2022). Guilt aversion in (new) games: Does partners' vulnerability matter? *Available at SSRN.*

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.

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46.

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*, 60(3):833–882.

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., 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.

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.

Charness, G., Samek, A., and van de Ven, J. (2022). What is considered deception in experimental economics? *Experimental Economics*, 25(2):385–412.

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.

d'Adda, G., Dufwenberg, M., Passarelli, F., and Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, 124:288–304.

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.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2023). Promises or agreements? moral commitments in bilateral communication. *Economics Letters*, 222:110931.

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 Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30(2):163–182.

Dufwenberg, M. and Nordblom, K. (2022). Tax evasion with a conscience. *Journal of Public Economic Theory*, 24(1):5–29.

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.

Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness—intentions matter. *Games and Economic Behavior*, 62(1):287–303.

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 Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.

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.

Ortmann, A., Fitzgerald, J., and Boeing, C. (2000). Trust, reciprocity, and social history: A re-examination. *Experimental Economics*, 3:81–100.

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.

Appendix A Literature

Table A.1 presents a list of published papers, citing Battigalli and Dufwenberg (2007) with an explicit reference to guilt aversion as a motivation of behavior and including an experiment.[31] The list includes two pioneer experiments published before the theoretical foundations of Battigalli and Dufwenberg (2007), such as Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2006), which currently are the most cited studies. We interpreted them as forerunners of Battigalli and Dufwenberg (2007), as they test the traditional theory of guilt aversion applied to certain games. This list shows that 43.18% of the papers focus on Trust games and 38.64% on Dictator games. It means that only 22.73% of the literature on guilt aversion has investigated other games.

Article	Game
Dufwenberg and Gneezy (2000)	Lost Wallet
Charness and Dufwenberg (2006)	Trust
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)	Trust
Dufwenberg et al. (2011)	Coordination
Pelligra (2011)	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
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)	Dictator
Inderst et al. (2019)	Trust
Morell (2019)	Dictator
Ciccarone et al. (2020)	Dictator
d'Adda et al. (2020)[32]	Dictator
Ghidoni and Ploner (2020)	Lost Wallet
Pelligra et al. (2020)	Trust
Danilov et al. (2021)	Dictator
Peeters and Vorsatz (2021)	Prisoner Dilemma
Dufwenberg and Nordblom (2022)	Tax compliance
Di Bartolomeo et al. (2023)	Dictator

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

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

[32] Although the motivation resembles guilt, they "consider players' expectations regarding how one ought to behave, rather than regarding how one will actually behave." (Battigalli and Dufwenberg, 2022, p. 857)

Appendix B 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*.[33] 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*.

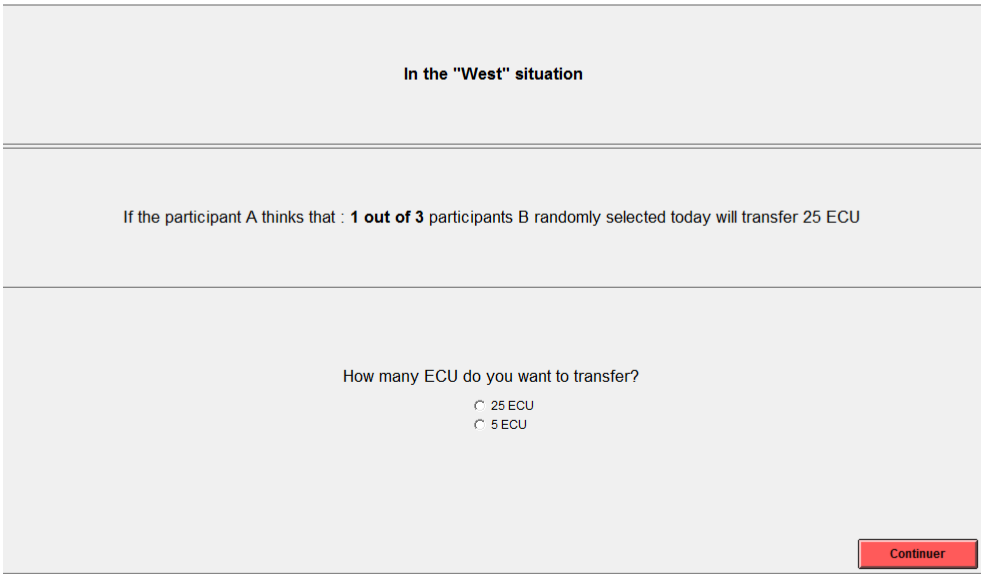

Figure B.1: Screenshot in Treatment A

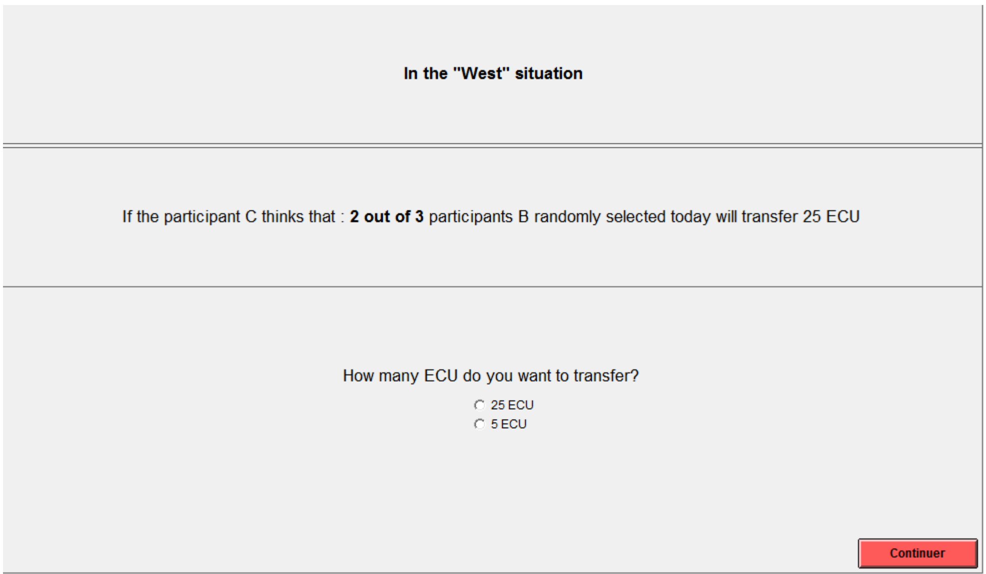

Figure B.2: Screenshot in Treatment C

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.

[33]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 ————————————

Appendix C A-subjects' Results

The choice of *In* by the 96 A-subjects of our study varies considerably across games. Pooling the two treatments, *In* is chosen by 48.87% of the A-subjects in the Investment game, 75.00% in the Exploitation game, 20.83% in the Donation game, and 70.83% in the Reversed-Investment game (Table C.1). We reject the null hypothesis that the proportion of A-subjects choosing *In* is the same across games (Cochran Q test; p-value = 0.000). Consistently, pairwise comparisons show that this proportion is significantly different across games (McNemar tests; highest p-value = 0.001 for Investment *vs.* Reversed-Investment game), except when we compare the Exploitation and the Reversed-Investment games (75.00% *vs.* 70.83%; p-value = 0.584).

Table C.1: Proportion of *In* choices across games, by first-order belief

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

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

We also estimate separate Logit regressions for each game with the choice of *In* as the dependent variable, and A-subjects' first-order beliefs and SVO angle as the main independent variables (Table C.2). 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 Exploitation and Reversed-Investment games. Moreover, this frequency of *In* choices also increases significantly in their prosociality (SVO angle) in the Investment and Donation games, but not significantly so in the Exploitation and Reversed-Investment games.

Table C.2: Likelihood of A-subjects choosing *In*, by game

	Investment		Exploitation		Donation		Rev-Investment	
FOB	0.491***	0.397**	-0.103	-0.172	0.422***	0.463***	-0.062	-0.125
	(0.147)	(0.156)	(0.138)	(0.138)	(0.084)	(0.103)	(0.143)	(0.141)
SVO angle	0.011***	0.009***	-0.004	-0.004	0.007***	0.007***	0.006	0.005
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.004)	(0.003)
Treatment A	0.160*	0.197**	-0.050	-0.020	-0.026	-0.003	0.095	0.069
	(0.089)	(0.090)	(0.086)	(0.086)	(0.071)	(0.070)	(0.094)	(0.093)
Order	0.011	0.036	0.085*	0.078	0.006	0.024	-0.049	-0.042
	(0.052)	(0.050)	(0.045)	(0.048)	(0.028)	(0.027)	(0.038)	(0.038)
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	-50.963	-46.122	-33.255	-29.341	-56.075	-50.634
Prob>chi2	0.000	0.000	0.196	0.107	0.000	0.000	0.441	0.146
Pseudo R2	0.181	0.273	0.056	0.145	0.323	0.403	0.032	0.126

Notes: Table C.2 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses. “FOB” for first-order beliefs α_{AB}. “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.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix D B-subjects' 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 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 followed 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.

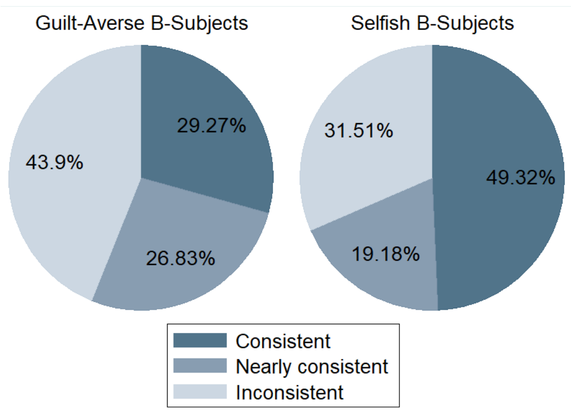

Figure D.1: Distribution of B-subjects' consistency of behavior

D.3 Likelihood of B-subjects choosing *Right* with induced or stated SOB

Table D.2: Likelihood of B-subjects choosing *Right* with stated or induced SOB, by game

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

Notes: Table D.2 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses. "Induced SOB" corresponds to the four β_{Bj} presented to B-subjects when making their choice of *Right* or *Left*. "Stated SOB" 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 Sensitivity to guilt measured by the switching SOB

Figure D.2 shows the distribution of switching SOB in each game-treatment combination for B-subjects consistent with Eq. (3). Keeping the game constant, the distributions of switching SOB do not differ significantly across treatments, except for the Donation game (Kruskal-Wallis test, p-value = 0.029).

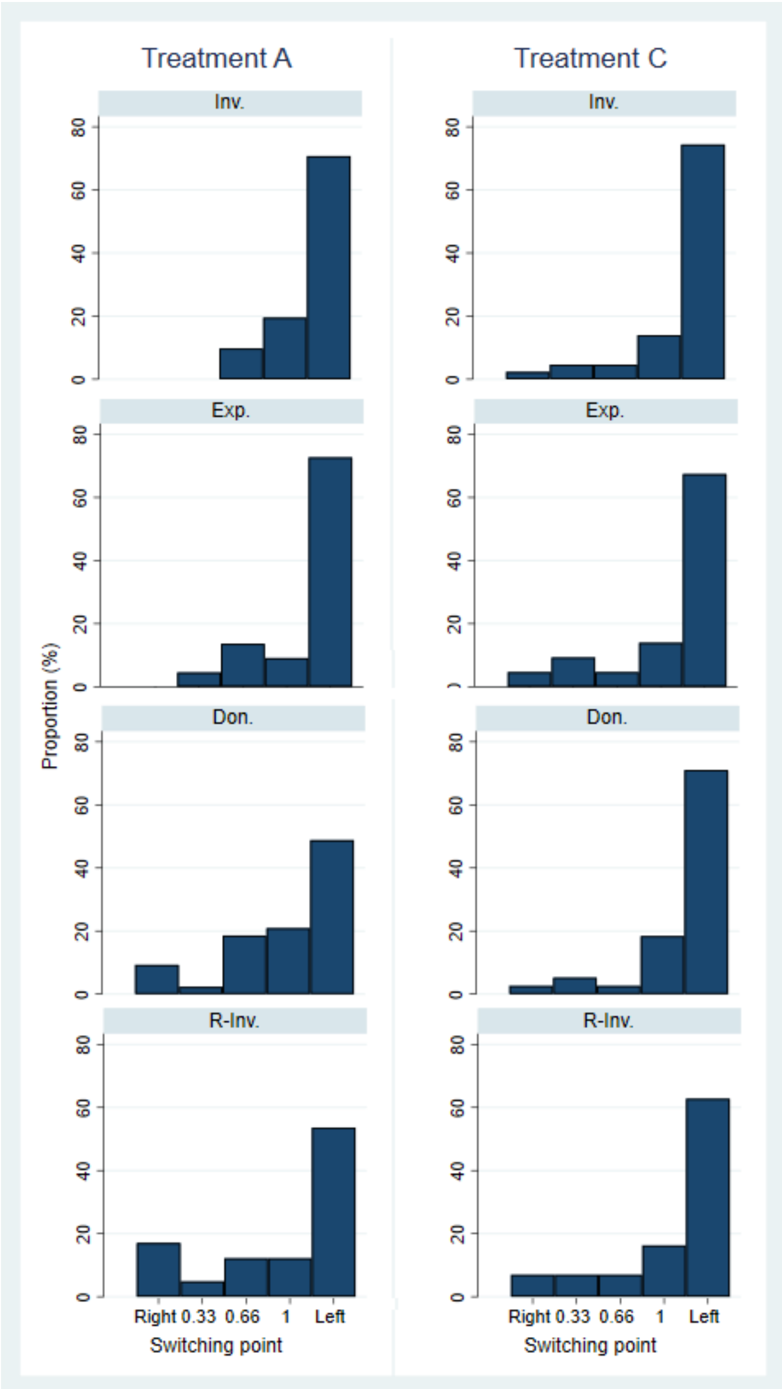

Figure D.2: Distribution of B-subjects' switching second-order beliefs
The figure reads as follows. In the Reversed-Investment game under Treatment A, 17.07% of the B-subjects always chose *Right*, 4.88% had a switching SOB from *Left* to *Right* of 0.33, *i.e.*, they chose *Left* for an induced SOB equal to 0 and *Right* for an induced SOB in in {0.33, 0.66, 1}, and so on.

D.5 Likelihood of B-subjects choosing *Right* depending on co-players' vulnerability or decision rights

Table D.3: Likelihood of B-subjects choosing *Right* depending on co-players' characteristics

	(1)	(2)	(3)	(4)
Induced SOB	0.188***	0.139***	0.188***	0.176***
	(0.024)	(0.043)	(0.024)	(0.035)
Payoff vulnerable	-0.073***	-0.074*	-0.073***	-0.074***
	(0.024)	(0.043)	(0.024)	(0.024)
Endowment vulnerable	0.010	-0.054	0.010	0.010
	(0.023)	(0.042)	(0.023)	(0.023)
Both vulnerable	0.013	0.033	0.013	0.013
	(0.033)	(0.061)	(0.033)	(0.033)
Treatment A	0.069*	0.068*	0.069*	0.055
	(0.037)	(0.037)	(0.037)	(0.046)
Payoff vul. * Ind. SOB	-	0.003	-	-
		(0.063)		
Endowment vul. * Ind. SOB	-	0.113*	-	-
		(0.062)		
Both vul. * Ind. SOB	-	-0.038	-	-
		(0.090)		
Treatment A * Ind. SOB	-	-	-	0.023
				(0.046)
Stated SOB	0.166***	0.165***	0.166***	0.166***
	(0.040)	(0.040)	(0.040)	(0.040)
Reported Guilt	0.017**	0.017**	0.017**	0.017**
	(0.007)	(0.007)	(0.007)	(0.007)
Order	-0.013*	-0.013*	-0.013*	-0.013*
	(0.008)	(0.008)	(0.008)	(0.008)
Personality	Yes	Yes	Yes	Yes
Demographics	Yes	Yes	Yes	Yes
N Observations	1536	1536	15636	1536
N Subjects	96	96	96	96
Log-likelihood	-593.449	-591.041	-593.449	-593.325
Wald Chi2	123.25	126.06	123.25	123.56

Notes: Table D.3 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses."Induced SOB" corresponds to the four β_{Bj} presented to B-subjects when making their choice of *Right* or *Left*. "Payoff vulnerable", "Endowment vulnerable" and "Both (payoff and endowment) vulnerable" refer to the dimensions of vulnerability of A-subjects and C-subjects (see Figures 1-4). "Stated SOB" corresponds to the second-order beliefs reported by the B-subjects in the belief elicitation stage. "Reported Guilt" takes value between 0 and 10. "Order" is the rank order of the game, from 1 to 4. "Personality" controls correspond to the subjects' self-reported prosociality, risk aversion and patience. "Demographics" controls include age, gender, frequency of past participation in experiments, majoring in business. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D.6 Structural estimation including inequity-averse B-subjects

Table D.4 reports the structural estimates of the mean guilt sensitivity in each game-treatment combination, considering only B-subjects with the behavioral patterns consistent with Eq. (3) (*i.e.*, choosing always *Left* or always *Right* regardless of the four α_{jB}, or switching from *Left* to *Right* as α_{jB} increases), which represent, on average, 87.54% of the B-subjects.

Table D.4: Structural estimates of guilt sensitivity for B-subjects disclosing behavior consistent with Eq. (3)

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

D.7 Follow-Up Study

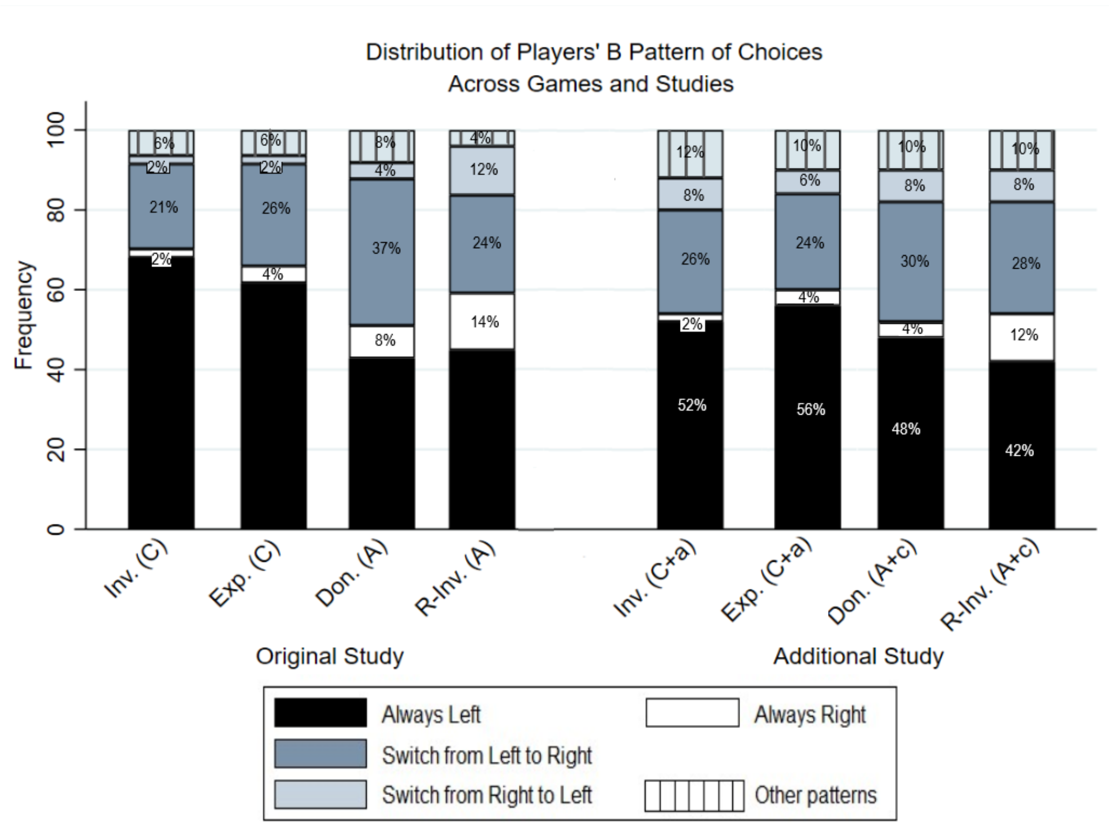

Figure D.3: Distribution of B-subjects' pattern of choices across games and studies

Capital letters indicate the co-player on whom expectations B-subjects can condition their decision. Small letters indicate the co-player whose expectation is directly given to B-subjects.