Guilt and Fairness*

Alessandro Stringhi[†1]

[1]*Università di Siena, Siena, Italy*

July 10, 2024

Abstract

Guilt Aversion and Inequity Aversion are pivotal concepts in understanding human behavior in situations involving trust dynamics. Inequity Aversion explains trustworthiness through a preference for fairness, as trustworthiness leads to more equitable distributions. Conversely, Guilt Aversion posits that people act to avoid the guilt associated with betraying others. As both preferences provide justifications for trustworthiness, distinguishing between them solely through observed behavior poses a significant challenge. In this work, I aim to disentangle the effects of Guilt and Inequity Aversion to identify the main driver of pro-social behavior in a theory-driven experiment based on the trust game. I show theoretically that by increasing the stakes for the trustor, the two preferences have opposite predictions on trustworthiness. The experimental design, informed by this theory, features a doubling of the trustor's payoffs. The results indicate that the preferences for equality are the main determinants of trustworthiness.

JEL classification: A13; C72; C91; D01; D91

Keywords: Inequity aversion, guilt aversion, beliefs, trust

1 Introduction

Trust is the foundation of every relationship, whether between spouses (Fehr 1988, Dufwenberg 2002), between employers and employees (Kreps et al. 1990, Dirks and Ferrin 2002), and it is the societal glue that leads to better economic outcomes (Arrow 1972, Fukuyama 1996, Knack and Keefer 1997). In order to understand the basis for trustworthiness, and consequently, trust, we must depart from the notion of selfish decision-makers (*homo oeconomicus*) and start considering other-regarding preferences. In this framework, two prominent other-regarding preferences, Inequity Aversion (Fehr and Schmidt 1999) and Guilt Aversion (Battigalli and Dufwenberg 2007), help us understand the basis of trustworthiness. Both cited preferences offer behavioral justifications for trustworthiness (Ciriolo 2007, Della Lena et al. 2023). The selfish behavior of the trustee results in disutility for themselves, with each preference providing a different explanation for this. In this work, I exploit the different motivations at the root of the two preferences. I compare the two and derive theoretical predictions that will inform the design of an experiment able to test the

*I gratefully acknowledge funding from the Italian Ministry of Education Progetti di Rilevante Interesse Nazionale (PRIN) grant 2017ELHNNJ and from Region Tuscany grant Spin.Ge.Vac.S. I thank the GREDEG research group and Giuseppe Attanasi for their support in the early stages of this work. I wish to thank Marco Stimolo, Pierpaolo Parrotta and my advisors, Pierpaolo Battigalli and Paolo Pin for their helpful comments.

[†]alessandro.stringhi@unisi.it

two preferences against each other.

For the purposes of this discussion, I focus on situations where the trustee's self-centered actions result in financial gain for themselves at a significant loss for the trustor, leading to an unfair distribution of assets or earnings. These situations closely relate to moral hazard, making them highly relevant to economics. The essence of these situations is perfectly captured and summarized by the Trust Games, even more so by the Trust Minigames employed in this work. The Fehr and Schmidt Inequity Aversion model (Fehr and Schmidt 1999) justifies trustworthiness by leveraging the preference for equitable distributions of the trustees. In the typical Trust Game, the selfish behavior of the trustee results in an inequitable distribution of payoffs, benefiting the trustee. If the trustee is concerned about fairness and equality, they will dislike the inequality generated by their behavior, leading them to opt for pro-social behavior that results in a fairer distribution of payoffs. This concern for equity potentially makes the trustee trustworthy, which is recognized by their counterpart, leading to trust. The Battigalli and Dufwenberg Guilt Aversion model (Battigalli and Dufwenberg 2007) is rooted in a different feeling, guilt. The model postulates that someone feels guilty when they disappoint others. To avoid this negative feeling, a person adjusts their behavior accordingly. In the context of the Trust Game, selfish behavior causes an unexpected loss for the trustor, leading to disappointment, which the trustee seeks to avoid by adjusting their behavior. Trustworthiness emerges as a consequence of the trustee's desire to avoid the guilt associated with disappointing the trustor through selfish behavior. While the former ascribes the disutility to the inequality generated, the latter attributes it to the desire to avoid causing disappointment to another person. Despite the profoundly different motivations underlying these preferences, it is normally challenging to discern between them during an experiment. This is because both preferences can explain and justify pro-social behavior in the laboratory, and it is not possible to infer motivations solely from observing behavior. The primary difference leveraged in the experimental literature to find evidence in support of either preference is the correlation, or lack thereof, between beliefs and behavior. Guilt Aversion, being a belief-dependent preference, displays a positive correlation between the trustee's behavior and their second-order beliefs, namely the beliefs about the trustor's beliefs about their behavior. Conversely, Inequity Aversion, in its mathematical formulation, is silent about any correlation between behavior and second-order beliefs. This correlation was used in Charness and Dufwenberg (2006) as evidence for Guilt Aversion. These findings were later challenged by Vanberg (2008), and in this work, I aim to continue the discussion. The main novelty of this work is that I do not leverage the correlation between beliefs and behavior, but instead I modify the trustor's material payoff. This manipulation allows me to test the two preferences directly, as it has a direct effect on the trustee's preferences without affecting their material payoffs, and this effect can be either positive or negative depending on which preference is more relevant for them.

Trust is also closely related to inequality (Rothstein and Uslaner 2005), and in many real-life situations, one party often has disproportionately greater stakes, typically the trustor. Interestingly, the two preferences react differently to increasing inequality between the two parties, the crucial aspect I exploit in this work. For example, consider a large corporation concerned about employee misbehavior. If society strongly disapproves of inequality, especially disadvantageous inequality, the likelihood of misbehavior is very high given the enormous profits of the corporation compared to the salary of the employees. In contrast, in a society highly sensitive to guilt, the likelihood of misbehavior would be close to zero due to the high stakes involved. Conversely, a small company or shop would have the opposite concerns. With fewer disparities in wealth and status between employees and the owner, there may be less envy stemming from the aver-

sion for inequality. However, in a society highly sensitive to guilt, the likelihood of misbehavior within a small company may paradoxically increase. Due to the lower stakes involved in smaller businesses, employees may perceive the consequences of their actions as less severe, leading to a higher propensity for misconduct. In this light, it becomes greatly important for economists to gain a clear understanding of the implications of behavioral models such as the ones under investigation. This is even more relevant given the growing interest in the role of inequality in our society, which affects the interpersonal dynamics at play in principal-agent problems involving disparities between the two parties.

In light of the interpersonal dynamics in principal-agent problems and the growing interest in the role of inequality, a significant question arises: which preference is more relevant when deciding whether or not to repay trust? This question remains unanswered, especially regarding the identification of Guilt Aversion, as highlighted in Fehr and Charness (2023). This study succeeds in providing a clear-cut answer to this research question. Through a controlled laboratory experiment involving the manipulation of trustor payoffs in a Trust Minigame, I am able to disentangle the behavioral implications of Guilt and Inequity Aversion. The experimental results document that subjects are predominantly concerned with inequality. These results have a more general validity and are not confined to individuals' behaviors in the Trust Game alone. The choice of using this specific strategic interaction is instrumental in depicting the conflict between these two kinds of preferences and in neatly disentangling their behavioral implications. However, these results can also hold for other games on social dilemmas, where both preferences could be relevant but would not allow for a clear disentangling or could be obscured by the presence of other dynamics.

This paper is inspired by the discussion stemming from Vanberg (2008) in response to the findings presented in Charness and Dufwenberg (2006). In their original paper, Charness and Dufwenberg (hereafter referred to as C&D) showed that non-binding pre-play communication enhanced trust and cooperation in a Trust Minigame. They attributed this effect to the promises made by the trustee, which raised the expectations of the trustor and, consequently, increased the guilt of the trustee if the promises were broken. All the findings of the paper are consistent with the predictions of Guilt Aversion, which the authors utilized as evidence in its favor. The primary evidence is the observed increase in cooperation when there is communication between the two parties. Promises contribute to heightened beliefs and, consequently, increase the potential sense of guilt. This increased cooperation follows an increase in both first- and second-order beliefs. Moreover, the authors find a positive correlation between second-order beliefs and the behavior of the trustor, a correlation which is predicted by Guilt Aversion but not by other preferences that are not belief-dependent. Subsequently, Vanberg challenged the original interpretation by arguing that C&D's results might primarily reflect preferences for promise-keeping. Vanberg designed an experiment capable of distinguishing and testing two alternative explanations for promise-keeping: *expectation-based* and *commitment-based*. The former explanation aligns with C&D's interpretation, and it is consistent with Guilt Aversion. The latter explanation aligns with models that assume a fixed cost of lying and relies on Inequity Aversion (Ellingsen and Johannesson 2004). The results of the experiment, which featured a design allowing for the independent variation of promises and beliefs, suggest that people have a preference for promise-keeping per se, challenging the results of Charness and Dufwenberg.

This discussion is still ongoing, with several papers presenting mixed evidence on the topic. In Ederer and Stremitzer (2017), the authors investigated the effect of exogenous variation of second-order beliefs. Their evidence shows the promisor's aversion to disappointing a promisee's ex-

pectation, and they proposed a model of conditional Guilt Aversion to explain their findings. Additional findings in support of Guilt Aversion, or more precisely in support of the *expectation-based explanation* for promise-keeping, come from Di Bartolomeo et al. (2023). This new evidence is in contrast with previous evidence found by the same authors in (Di Bartolomeo et al. 2019), which supports the *commitment-based explanation* instead. The relevance of the topic, the joint testing of Guilt and Inequity Aversion, is not just relegated to the discussion mentioned. Contrary to the Fehr and Schmidt model of inequity aversion that can be easily used to explain a wide variety of behavior observed in the lab, the Guilt Aversion model is more subtle to test. This is due to the nature of the model being belief-dependent. Other works have tried to test for Guilt Aversion (Ellingsen et al. 2010, Khalmetski 2016, Bellemare et al. 2017) by employing the intrinsic correlation between beliefs and behavior, finding positive evidence. In contrast to all the previously cited works that tested for Guilt Aversion, my treatment manipulation does not influence behavior through an exogenous variation of beliefs, but it acts on the other-regarding aspect of the model. This allows for an easier comparison between the two models, avoiding the criticisms raised by Vanberg (2008).

My work develops alongside the previously mentioned studies, but it follows a different and more direct approach than its predecessors. The main focus in the literature so far has been on the role of exogenous variation of beliefs on behavior, often linking them to the two explanations for promise-keeping, *expectation-based* and *commitment-based*, which are associated with Guilt and Inequity Aversion, respectively. Departing from this approach, I aim to directly test the two preferences against each other by influencing other-regarding preferences and only secondarily, beliefs. To achieve this, I develop an experiment in which I double the payoffs of the trustor in a Trust Minigame for the treatment group, while the Trust Minigame for the control group closely resembles the one employed in Charness and Dufwenberg (2006). The payoffs of the trustee are left unchanged between treatments. This payoff manipulation serves as the mechanism to disentangle the two preferences, as it triggers different reactions according to each one. Doubling the trustors' payoffs leads to greater inequality between them and the trustees. This inequality is detrimental for the trustees, who consequently behave more selfishly. Conversely, the higher payoffs for the trustors are associated with higher expected payoffs for them. This heightened expectation increases the potential disappointment for the trustors in case the trustees behave selfishly. Therefore, if the trustees wish to avoid a sense of guilt, they will tend to behave more pro-socially. The behavior of the trustees, in turn, influences the trustors' behavior, as they anticipate the trustees' actions. Since the two theories predict opposite effects of the treatment, I am able to directly disentangle and test the two preferences against each other. Changing only the payoffs of the trustors, while keeping those of the trustees unchanged, is a crucial aspect of my design that allows me to disentangle the two types of preferences. This design choice ensures that any changes in the behavior of the trustees, which is the main focus of this work, are driven by changes in their other-regarding preferences as influenced by the manipulated payoffs of the trustors.

The results reveal a significant reduction in pro-social behavior in the treated group. Specifically, when comparing the control group with the treated group, it is observed that the proportion of subjects exhibiting pro-social behavior is halved. In the control group, 34.4% of trustors and 44.1% of trustees played pro-social actions, whereas in the treated group, these percentages decreased to 14.1% and 20.3% respectively for trustors and trustees. Similarly, both first- and second-order beliefs about the trustee's behavior decrease significantly. In the control group, the average reported first-order belief is 0,30, while in the treated group, it decreases to 0,20. Correspondingly, trustees' reported second-order beliefs decrease from an average of 0,35 in the control group to 0,20 in the

treated group. These findings are consistent with the predictions of the Inequity Aversion model, indicating that the subjects strongly reacted to disadvantageous inequality.

The experimental design is supported by an intuitive theoretical analysis. Using rationalizability, I solve the Trust Minigame employed in the experiment twice, once for each preference. This analysis reveals that the two theories exhibit opposing monotonic trends concerning the increase in the trustor's stakes. As the trustor's stakes increase, so do their expectations regarding material gains. In the event that the trustee fails to repay the trust, this leads to significant disappointment for the trustor. Consequently, a guilt-averse trustee seeks to avoid causing such disappointment, thus promoting more pro-social behavior. Conversely, with the growth of the trustor's stakes comes an increase in the inequality between the two parties involved. According to Inequity Aversion theory, as this inequality escalates, so does the envy experienced by the trustee, resulting in a reduction in pro-social behavior. Moreover, this analysis sheds light on the distinct behavioral implications of the two preferences in the presence of inequality between the parties involved.

This paper is organized as follows: In Section 2, I analyze the two models and demonstrate their differing behavior when the payoffs of the trustor are increased. This justifies the experimental design presented later in Section 3, while in Section 4 I outline the behavioral predictions derived from the theoretical analysis. In Section 5, I present the results of the experiment, and finally, Section 6 provides the conclusion.

2 Theoretical Analysis

The primary goal of this section is to introduce the treatment employed in the experiment and show that with this treatment I am able to disentangle the two preferences under investigation - Guilt and Inequity Aversion. In this section I solve the game twice, once for each preference. I use this approach, alongside other simplifying assumptions, in order to make the discussion as streamlined as possible. In this way I am able to show more clearly the effect on behavior and beliefs of the two different behavioral channels.

In behavioral economics, the Fehr & Schmidt Inequity Aversion (Fehr and Schmidt 1999) and the Battigalli & Dufwenberg Guilt Aversion (Battigalli and Dufwenberg 2007) are two prominent models that can explain a wide variety of behaviors observed in the lab, behaviors that are in contrast with the assumptions of rationality and selfishness inherent in the *homo oeconomicus* framework. Although both preferences correctly predict a positive fraction of pro-social outcomes in a Trust Minigame, their predictions vary when the first-mover's material payoffs are increased while the payoffs of the second-mover are held constant. In this section, I will show that when the initial allocation of the first-mover (trustor) is increased, becoming higher than the one of the second-mover (trustee), the Inequity Aversion model predicts a low frequency of cooperative outcomes. Instead, the Guilt Aversion model predicts the opposite, and the difference in the frequency of cooperative outcomes predicted by the two models grows as the trustor's payoffs increase. This result guides the design of the experiment, which is meant to test these theories against one another. To prove this first result, the two preferences are analyzed separately. I make the assumption, maintained throughout, that the trustor is selfish and risk-neutral, while the trustee is inequity (resp. guilt) averse.[1] The type of preference for the trustee, either Guilt or Inequity Aversion, is

[1]This assumption is known as role-dependent inequity (resp. guilt) aversion. It serves to clarify the results and make the behavioral mechanisms transparent and the intuition behind the main results of this section remain valid

assumed to be common knowledge. Despite knowing the type of preference of the co-player, the trustor doesn't know their level of inequity (resp. guilt) aversion; therefore, this is a game with incomplete information.

To begin, I introduce a parameterized version of the Trust Minigame, depicted in Figure 1. This parameterization allows me to illustrate the distinct monotonic properties of the two models in relation to the variable m. The parameter m rescales the wealth of the trustor, and it is the ratio between the trustor's and trustee's initial wealth. The game in Figure 1 simulates the investment decision-making process[2] between two individuals. For the sake of clarity, I will now refer to the first- and second-movers, respectively, as Ann and Bob. In this scenario, Bob presents Ann with an investment opportunity that carries the potential to double their initial wealth, mW and W respectively. However, it's important to note that their initial wealth is not equal. Ann possesses m times the initial wealth of Bob, W, where $m \geq 1$ represents a parameter that determines the ratio between Ann's and Bob's initial wealth. This parameter will serve as the treatment variable in the experiment and plays a pivotal role in this analysis. Ann is presented with a choice between *In*vesting in Bob's project or choosing not to and opting *Out*. If Ann decides to invest, Bob faces a subsequent decision. He can choose to fulfill his obligations, doubling the initial investment, and then *Share* the profit with Ann proportionally, based on each investor's initial contribution. Alternatively, Bob can choose to *Take*, all for himself, a reduced profit, equal to twice his initial investment plus an additional gain of gW, leaving Ann with nothing.

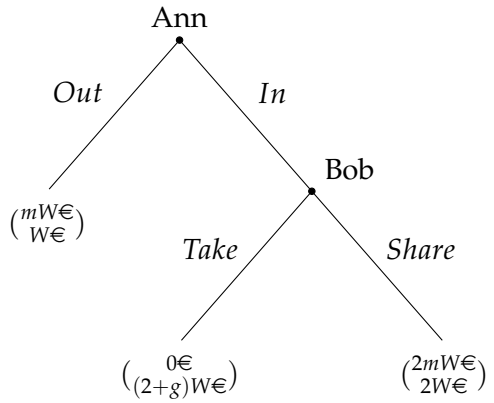


Figure 1: Parameterized Trust Minigame with Material Payoffs.

The behavior of Bob is contingent upon the variable m. As m increases, so does the disparity in wealth between Ann and Bob. If Bob exhibits Inequity Aversion, this increased inequality reduces his utility when choosing *Share*. Simultaneously, as m rises, Ann's expected payoff also increases, so does her disappointment when Bob choose *Take*. This will increase Bob's guilt if he is guilt-averse. As we can already see there is a tension between the two preferences under investigation. The manipulation of m and its direct effects on the two preferences is the main intuition behind the forthcoming theoretical analysis. In this analysis, I briefly introduce the two preferences. Then, I

even if both players are considered inequity (resp. guilt) averse. Attanasi et al. (2016) delves into the details of why, in the case of guilt, this assumption is justified by the literature in psychology and explores how the results change when this hypothesis is removed. While for the case of Inequity Aversion is possible to notice that the trustor's preferences over outcomes remain unaltered, $(In, Sh) \succ (Out) \succ (In, Tk)$. This is true for any level of aversion to inequality, and this suggests that her decision is largely based on her strategical reasoning about Bob's behavior.

[2]The Trust Game was originally referred to as the Investment Game (Berg et al. 1995).

proceed to solve the game twice using rationalizability, once for each preference, demonstrating how the best reply correspondence depends on the value of m.

Notice that the decision to solve the game twice, once and independently for each preference, while assuming common knowledge about them, is a simplification adopted exclusively for the sake of clarity and parsimony so as to clearly illustrate the behavioral implications of each preference. Doing so aligns better with the objectives of this paper, which aims to test experimentally Guilt Aversion against Inequity Aversion. For this purpose, providing a clear indication of each of the behavioral channels involved separately is more valuable than a more complex, albeit theoretically accurate, analysis that would largely rely on the same intuitions and behavioral motivations.

2.1 Guilt Aversion

Guilt is a form of emotional distress, and, using the definition provided by psychologists Baumeister et al. (1994), it arises when individuals cause harm, loss, or distress to a relationship partner. Those who feel guilty about hurting their partners by failing to meet their expectations often adjust their behavior to avoid experiencing this emotional distress. Building upon this definition, Battigalli and Dufwenberg (2007) formulated a game theoretical model of Guilt Aversion, delving into how this emotional state influences strategic interactions. This model uses the framework of psychological game theory, first developed by Geanakoplos et al. (1989) and extended in Battigalli and Dufwenberg (2009). In a psychological game, the utility of each player depends not only on the outcome of their actions but also on their beliefs, beliefs on what others believe, and so on. Often, these preferences are called belief-dependent preferences to highlight the presence of beliefs in the utility function.

In the formulation presented in Battigalli and Dufwenberg (2007), a typical player i experiences a disutility that is proportional to the disappointment of the other player j. Disappointment D_j is defined as the difference between the expected payoff and the actual received payoff, expressed formally as:

$$D_j[\pi_j; \alpha_j] = \max\left\{\mathbb{E}_{\alpha_j}[\tilde{\pi}_j] - \pi_j, 0\right\}, \tag{1}$$

where π_j is the material payoff j received, while $\mathbb{E}_{\alpha_j}[\tilde{\pi}_j]$ is the payoff she expected. This expectation is computed using j's subjective first-order belief about i's actions, α_j. Hence, the utility function of a guilt-averse player can be expressed as follows:

$$u_i(\pi_i, \pi_j, \alpha_j; \theta_i^G) = \pi_i - \theta_i^G \left(D_j[\pi_j; \alpha_j]\right), \tag{2}$$

where θ_i^G is i's sensitivity to guilt. Given that the first-order belief α_j is not directly observable by player i, player i must form a belief about j's first-order belief, leading to the second-order belief denoted by β_i.

For the purposes of this work, only some aspects of players' first- and second-order beliefs matter. In particular, I consider the subjective probability assigned by Ann to Bob playing *Share*, α_A^{Sh}. Similarly, I consider Bob's subjective (second-order) expectation of α_A^{Sh}, denoted by β_B^{Sh}. I also assume, consistent with Battigalli and Dufwenberg (2009), that players' behavior aligns with their plans, reflecting the same inference imposed by forward-induction reasoning. This choice further supports the focus on specific elements of the players' beliefs.

Type structure In order to model heterogeneity of traits and exogenous beliefs, I employ *Harsanyi types*. I assume that the set of types is a Cartesian product $T_r = \Theta_r^G \times \mathcal{E}_r$, where the set of epistemic types is $\mathcal{E}_r = [0,1]$, with $r \in \{A, B\}$ being the role of the player, trustor and trustee respectively. Thus, a type t_i of player i is a pair (θ_i^G, e_i), traits and epistemic types are assumed to be independent. The *epistemic type*, e_i, determines the beliefs about the co-player's type. Since the trustors are known to be selfish, $\Theta_A^G = \{0\}$, their type and their epistemic type coincide, $t_A = e_A$. While for the trustees $\Theta_B^G = [0, \overline{\theta}^G]$, with $\overline{\theta}^G$ being common knowledge.

Let $\mathcal{G}_\mathcal{E}$ be a family of cumulative distribution functions (CDF) $G_e(\theta^G) : \Theta_B^G \to [0,1]$,[3] with full support, differentiable and with continuous density function $g_e(\theta^G)$. Each of these CDF describes a possible distribution of guilt sensitivity parameters of the population. The real distribution of traits in the population of trustees is given by the CDF $G(\theta^G) \in \mathcal{G}_\mathcal{E}$. The real distribution is unknown, therefore each epistemic type e_i has a subjective belief about the real distribution, which is represented by $G_{e_i}(\theta^G) \in \mathcal{G}_\mathcal{E}$. Now, each epistemic type e_i parametrizes i's subjective probability about the fraction of the population with guilt sensitivity above a certain threshold $\hat{\theta}_p^G$, thus $e_i = 1 - G_{e_i}(\hat{\theta}_p^G)$. This parametrization is not unique, and it is determined by the choice of the threshold $\hat{\theta}_p^G$. This assumption will become clear later, when I will use two different thresholds, θ_{min}^G and θ_{max}^G. The distribution of the epistemic types is given by the CDF $H^p(e_i) : [0,1] \to [0,1]$, this true distribution is unknown to the players, thus each epistemic type e_i holds different beliefs about the type of the coplayer e_j. The beliefs of each e_i are given by a CDF $H_{e_i}^p(e_j)$. These distributions depend on the parametrization p used.

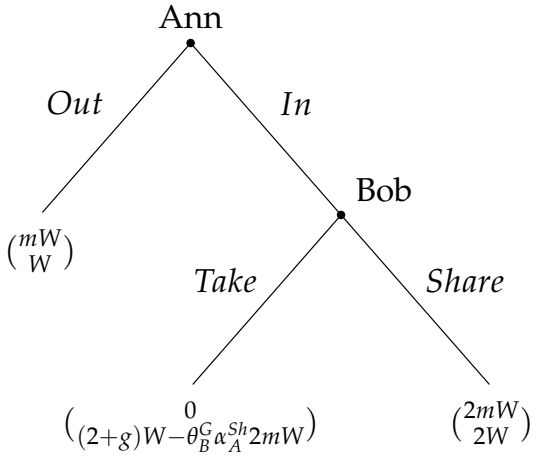


Figure 2: Trust Minigame with Psy-Utilities (Role-Dependent Guilt).

Best Reply Correspondence and Beliefs Now I will solve the game depicted in Figure 2 using rationalizability and elimination of non-best replies. Given Ann's rationality, she opts for *In* only if she believes that it will result in greater utility than *Out*, thus only if she trusts Bob, which means $\alpha_A^{Sh} \geq \frac{1}{2}$. Bob's choice depends on his sensitivity to guilt and second-order belief; he will play *Share* if $\theta_B^G \geq \hat{\theta}^G = \frac{g}{2m\beta_B^{Sh}}$. Since Bob believes in Ann's rationality and that she carries out her plan, he understands that she plays *In* if and only if $\alpha_A^{Sh} \geq \frac{1}{2}$, therefore any conditional second-order belief

[3] I drop the suffix because I focus exclusively on $G(\theta_B^G)$ since $\Theta_A = \{0\}$.

$\beta_B^{Sh} < \frac{1}{2}$ will violate the assumption of rationality and common belief of order one in rationality. Bob's best reply correspondence is depicted in Figure 3.

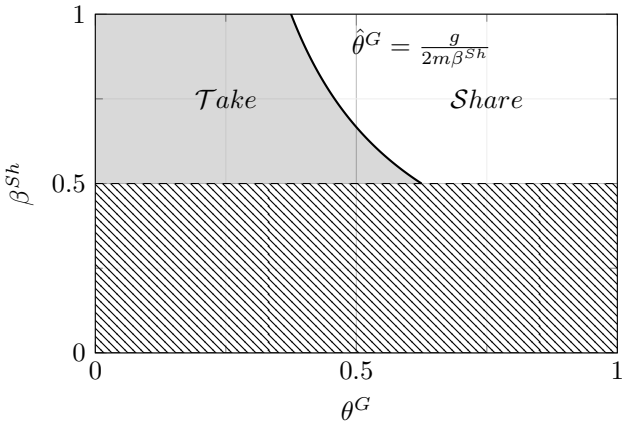


Figure 3: Bob's Best Reply Correspondence: Guilt Aversion

Monotonicity of the Best Reply and Beliefs The first result of this paper delves into how the likelihood of playing *Share* and *In*, as well as the beliefs, depend on the parameter m. Later, I will show how this dependence differs between Guilt and Inequity Aversion. Intuitively, the probability of Bob playing *Share*[4] is proportional to the white area of Figure 3. The same goes for the probability of playing of Ann playing *In* and the first and second-order beliefs, which are all determine by the probability of Bob playing *Share*. This area is delimited to the left by the threshold $\hat{\theta}^G = \frac{g}{2m\beta_B^{Sh}}$, and this threshold depends on the parameter m, and it is decreasing in it. Therefore, a higher value of m' compared to m will result in a larger area, as depicted in Figure 4, leading to a higher probability of Bob playing *Share*, and consequently Ann playing *In*, and higher first- and second-order beliefs.

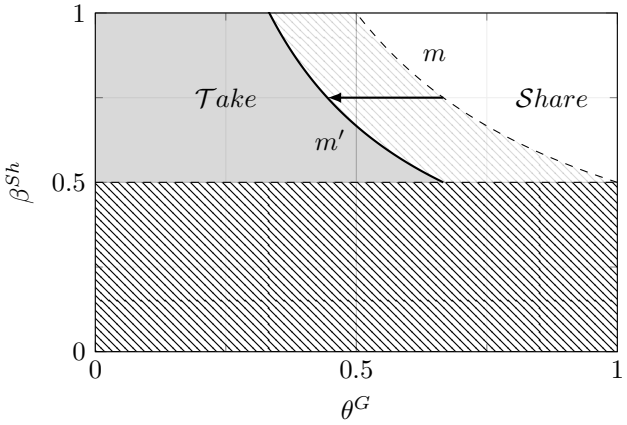


Figure 4: Monotonic Relationship in Bob's Best Reply: Guilt Aversion

Although intuitively every measure of interest should increase with m, this is not trivial to prove. This is due to the interdependence between actions and endogenous second-order beliefs. A common assumption that I also make is to assume that all CDF $G_e(\theta^G) \in \mathcal{G}_\mathcal{E}$, including the real

[4]Conversely, this can be viewed as the fraction of players in the population of trustees that play *Share*. In the experiment, this is what I will measure.

distribution $G(\theta^G)$, are independent from the game form, therefore independent from m. This assumption can be reasonable for exogenous traits like θ^G, and beliefs about said traits, but the same cannot be said about an endogenous variable and beliefs about endogenous variables like β^{Sh}. Therefore, even if the threshold decreases with m, I cannot ensure that under a different m' the distribution of second-order beliefs remains unchanged. Consequently, to avoid imposing constraints on the beliefs that might either dictate the result or contradict it, I need to make all the relevant variables depend only on exogenous variables and beliefs. This will lead to a weaker result that will not undermine the theoretical predictions for my experiment, as I make the design to choice to set m equal to 1 and 2 the control and in the treatment respectively. This design choice is discussed in more detail at the end of this section.

Rather than discussing the level of cooperation for each belief, I will focus on the minimum and maximum of those values, proving that they increase as m rises. The minimum and maximum values are determined by simply setting β^{Sh} to $\frac{1}{2}$ and 1, respectively, where $\frac{1}{2}$ represents the minimum value for β^{Sh} that remains consistent with rationality and common belief in rationality. Then, we can define two new thresholds: $\hat{\theta}^G_{min} = \frac{g}{m}$ and $\hat{\theta}^G_{max} = \frac{g}{2m}$. These new thresholds are independent from β^{Sh}, solving the problem.

Now, I can define the minimum and maximum probability of an individual drawn from the population of trustees playing *Share*, respectively. These are defined, respectively, as follows:

$$\begin{aligned} \min \mathbb{P}(Share) &= \int_{\frac{g}{m}}^{\overline{\theta}^G} g(\theta^G)\, d\theta^G = 1 - G(\frac{g}{m}), \\ \max \mathbb{P}(Share) &= \int_{\frac{g}{2m}}^{\overline{\theta}^G} g(\theta^G)\, d\theta^G = 1 - G(\frac{g}{2m}), \end{aligned} \tag{3}$$

where $G(\theta^G)$ is the true CDF that describes the distribution of guilt sensitivity in the population, which is unknown to both Ann and Bob.

From the pair of equations 3, I can define the first-order belief of a trustor. The minimum and maximum first-order beliefs are defined as follows:

$$\begin{aligned} \min \alpha_A^{Sh} &= \int_{\frac{g}{m}}^{\overline{\theta}^G} g_{e_A}(\theta^G)\, d\theta^G = 1 - G_{e_A}(\frac{g}{m}) = e_A^{min}, \\ \max \alpha_A^{Sh} &= \int_{\frac{g}{2m}}^{\overline{\theta}^G} g_{e_A}(\theta^G)\, d\theta^G = 1 - G_{e_A}(\frac{g}{2m}) = e_A^{max}, \end{aligned} \tag{4}$$

where the supersctipts of e_A^{min} and e_A^{max} indicate different parametrizations given by the different thresholds θ^G_{min} and θ^G_{max}. Then, the average minimum and maximum first-order beliefs are given by:

$$\begin{aligned} \mathbb{E}[\min \alpha_A^{Sh}] &= \int_0^1 e_A\, dH^{min}(e_A), \\ \mathbb{E}[\max \alpha_A^{Sh}] &= \int_0^1 e_A\, dH^{max}(e_A). \end{aligned} \tag{5}$$

An individual drawn from the population of trustors will choose *In* only if she holds the belief $\alpha_A^{Sh} \geq \frac{1}{2}$. The minimum and maximum probability that such an individual opts for *In* is deter-

mined by:

$$\begin{aligned}\min \mathbb{P}(In) &= \int_0^1 \mathbf{1}_{(\min \alpha_A^{Sh} \geq \frac{1}{2})}\, dH^{min}(e_A) = \int_0^1 \mathbf{1}_{(e_A^{min} \geq \frac{1}{2})}\, dH^{min}(e_A) = \int_{\frac{1}{2}}^1 dH^{min}(e_A),\\ \max \mathbb{P}(In) &= \int_0^1 \mathbf{1}_{(\max \alpha_A^{Sh} \geq \frac{1}{2})}\, dH^{max}(e_A) = \int_0^1 \mathbf{1}_{(e_A^{max} \geq \frac{1}{2})}\, dH^{max}(e_A) = \int_{\frac{1}{2}}^1 dH^{max}(e_A).\end{aligned} \tag{6}$$

The unconditional second-order beliefs, denoted by $\beta_B^{\varnothing}$ to distinguish them from β_B^{Sh}, the second-order beliefs conditional on Ann playing *In*, are defined as Bob's expectation of Ann's first-order beliefs. Therefore, they are determined by the following formulas:

$$\begin{aligned}\min \beta_B^{\varnothing} &= \mathbb{E}_{e_B}[\min \alpha_A^{Sh}] = \int_0^1 e_A\, dH_{e_B}^{min}(e_A),\\ \max \beta_B^{\varnothing} &= \mathbb{E}_{e_B}[\max \alpha_A^{Sh}] = \int_0^1 e_A\, dH_{e_B}^{max}(e_A),\end{aligned} \tag{7}$$

and their average values are given by:

$$\begin{aligned}\mathbb{E}[\min \beta_B^{\varnothing}] &= \int_0^1 \beta_B^{\varnothing}\, dH^{min}(e_B),\\ \mathbb{E}[\max \beta_B^{\varnothing}] &= \int_0^1 \beta_B^{\varnothing}\, dH^{max}(e_B).\end{aligned} \tag{8}$$

Notice that since $\beta_B^{\varnothing}$ are unconditional beliefs, they are not imposed to be greater than $1/2$.
Due to the monotonic properties of integral and cumulative distribution functions, it is straightforward to see that $\min \mathbb{P}(Share)|_{m'} \geq \min \mathbb{P}(Share)|_m$ for every $m' > m$. The same applies to $\max \mathbb{P}(Share)$. Since all other measures are derived from $\min \mathbb{P}(Share)|_m$ and $\max \mathbb{P}(Share)|_m$, they also follow that same trend. This leads to the first result:

Proposition 1 *Consider two Trust Minigames with parameters m and m' and populations of guilt-averse trustees described by $G(\theta^G)$ and $H^p(e_B)$ and trustors described by $H^p(e_A)$. Then, for every $m' > m \geq 1$, the following relationships hold:*

- $\min \mathbb{P}(Sh)|_{m'} \geq \min \mathbb{P}(Sh)|_m$, *and* $\max \mathbb{P}(Sh)|_{m'} \geq \max \mathbb{P}(Sh)|_m$;
- $\min \mathbb{P}(In)|_{m'} \geq \min \mathbb{P}(In)|_m$, *and* $\max \mathbb{P}(In)|_{m'} \geq \max \mathbb{P}(In)|_m$;
- $\mathbb{E}[\min \alpha_A^{Sh}]|_{m'} \geq \mathbb{E}[\min \alpha_A^{Sh}]|_m$, *and* $\mathbb{E}[\max \alpha_A^{Sh}]|_{m'} \geq \mathbb{E}[\max \alpha_A^{Sh}]|_m$;
- $\mathbb{E}[\min \beta_B^{\varnothing}]|_{m'} \geq \mathbb{E}[\min \beta_B^{\varnothing}]|_m$, *and* $\mathbb{E}[\max \beta_B^{\varnothing}]|_{m'} \geq \mathbb{E}[\max \beta_B^{\varnothing}]|_m$.

This first result shows how, under the assumption of guilt averse preferences of the trustee, the level of cooperation and trust, as well as the associated beliefs, vary with changes in the parameter m that determines the wealth ratio between players in the Trust Minigame. As m increases, the likelihood of playing *Share*, the likelihood of playing *In*, and the first- and second-order beliefs all exhibit a consistent pattern of increasing, reflecting a greater propensity for cooperation and trust. This result is primarily driven by the higher stakes for the trustor, leading to increased levels of disappointment and guilt if the trustee chooses to play *Share*.

It is essential to note, especially with regard to the experimental design, that for $m = 1$ and $m = 2$, we have $\theta_{min}^G|_{m=2} = \theta_{max}^G|_{m=1}$. This observation allows me to confidently assert that

$\mathbb{P}(Share)|_{m=2} \geq \mathbb{P}(Share)|_{m=1}$, which means that the minimum amount of *Share* in the treatment ($m = 2$) is greater than the maximum amount in the control ($m = 1$). This is regardless of the beliefs of the trustee. This is crucial because it shows that the assumption of Guilt Aversion being common knowledge, an assumption made exclusively for the sake of clarity, can be dropped without undermining the disentangling power of my treatment manipulation.

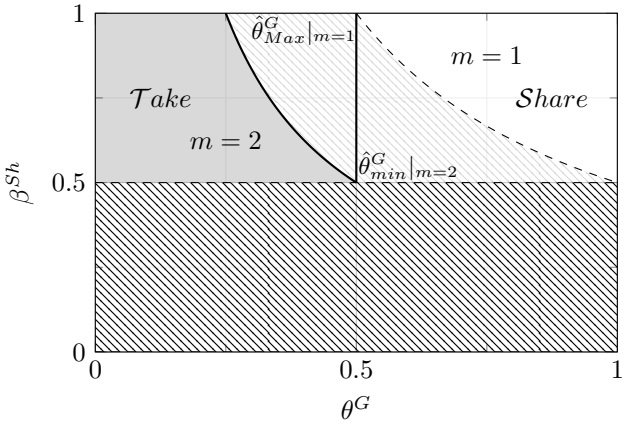


Figure 5: Thresholds: $m = 1$ and $m = 2$.

2.2 Inequity Aversion

The Inequity Aversion model, presented by Fehr and Schmidt (1999), explains how fairness impacts people's behavior. It suggests that individuals naturally dislike unfair situations and tend to adjust their actions to restore fairness. Inequity aversion is an other-regarding preference, meaning that the payoff of the co-player enters the utility function, a feature share with Guilt Aversion. Under Fehr and Schmidt (1999) formulation, Inequity Aversion states that a typical player i is willing to sacrifice part of their material payoff for a reduction in inequality between themselves and their counterpart j. The utility function of an inequity-averse player i is represented by the following formula:

$$u_i(\pi_i, \pi_j; \theta_i^I, \theta_i^S) = \pi_i - \theta_i^I \max\{\pi_j - \pi_i, 0\} - \theta_i^S \max\{\pi_i - \pi_j, 0\}. \tag{9}$$

The parameters $\theta_i^I \in [0, 1)$ and $\theta_i^S \in [0, 1)$ measure player i's sensitivity to disadvantageous inequality and advantageous inequality, respectively, which capture i dislike to get less (*I*nferiority) or more (*S*uperiority) than j. A common assumption is $\theta^I \geq \theta^S$, which implies that individuals dislike inequality more when it harms them.

Type structure Once again, I employ the *Harsanyi types* as did previously for Guilt Aversion. The set of types is the Cartesian product $T_r = \Theta_r^F \times \mathcal{E}_r$, with the set of traits being $\Theta_r^F = \Theta_r^I \times \Theta_r^S$ and $\mathcal{E}_r = [0, 1]$ being the set of *epistemic types*. Also for the case of Inequity Aversion, the trustors are known to be selfish, $\Theta_A^F = \{(0, 0)\}$, and their type and epistemic type coincide, $t_a = e_a$. Let $\mathcal{F}_\mathcal{E}$ be a family of cumulative distributions $F_e(\theta^I, \theta^S) : \Theta_B^F \to [0, 1]$, with full support, differentiable with continuous density function $f_e(\theta^I, \theta^S)$. The real distribution of traits in the population of trustees is given by the CDF $F(\theta^I, \theta^S) \in \mathcal{F}_\mathcal{E}$. The real distribution is unknown, therefore each epistemic type e_i has a subjective belief about the real distribution, which is represented by $F_{e_i}(\theta^I, \theta^S) \in \mathcal{F}_\mathcal{E}$. Now, each epistemic type e_i parametrizes i's subjective probability about the fraction of the population

with sensitivity to Superiority lower than a certain threshold $\hat{\theta}^S$, thus $e_i = 1 - \text{marg}_{\Theta^I} F_{e_i}(\hat{\theta}^S)$. The distribution of the epistemic types is given by the CDF $K(e_i) : [0,1] \to [0,1]$, which is unknown to the players. Each epistemic type e_i holds different beliefs about the coplayer's type e_j. The beliefs of each e_i is given by a CDF $K_{e_i}(e_j)$.

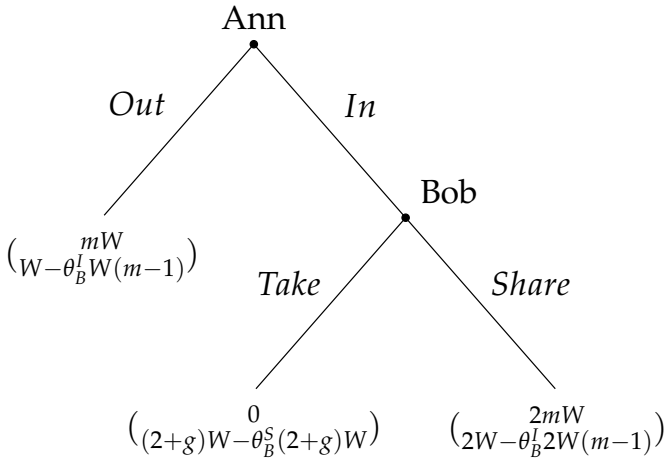


Figure 6: Trust Minigame with Utilities (Role-Dependent Inequity Aversion).

Best Reply Correspondence As I did previously, I will solve the game depicted in Figure 6 using rationalizability and elimination of non-best replies. Once again, Ann's choice will depend on whether she trusts Bob or not. Thus, given her rationality, she plays *In* only if she believes that Bob will play *Share* with high enough probability, $\alpha_A^{Sh} \geq \frac{1}{2}$. Unlike in the case of Guilt Aversion, Bob choice depends exclusively on his sensitivity parameters, θ_B^I and θ_B^S, and not on his second-order belief. He will play *Share* if and only if his sensitivity to advantageous inequality is high enough, namely $\theta_B^S \geq \hat{\theta}^S = \frac{g+\theta_B^I(m-1)}{2+g}$. Although not essential to Bob's best reply, Bob's second-order beliefs have the same constraint as they did in the case of Guilt Aversion. In other words, any beliefs $\beta_B^{Sh} < \frac{1}{2}$ will violate the assumption rationality and common belief of order one in rationality . Bob's best reply correspondence is depicted in Figure 7.

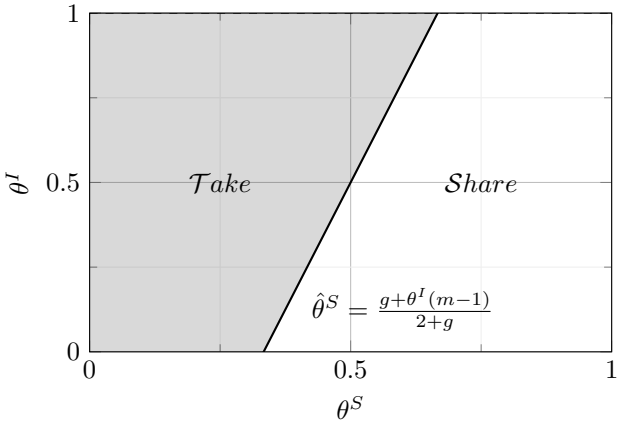


Figure 7: Bob's Best Reply Correspondence: Inequity Aversion

Monotonicity of the Best Replies and Beliefs Similar to Guilt Aversion, the probabilities associated with playing *Share*, *In*, as well as the first and second-order beliefs, are directly correlated with the white area depicted in Figure 7, which is determined by the threshold $\hat{\theta}_B^S = \frac{g+\theta^I(m-1)}{2+g}$. Unlike the previous case, the threshold $\hat{\theta}^S$ increases with m. Therefore, a higher value of m' compared to m results in a smaller white area, as depicted in Figure 8.

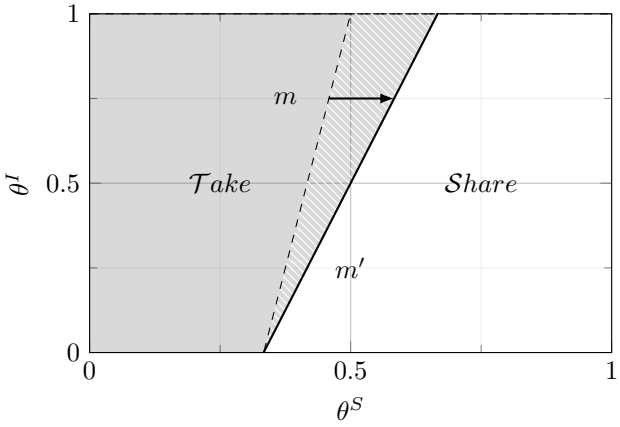


Figure 8: Monotonic Relationship in Bob's Best Reply: Inequity Aversion

Since the Inequity Aversion utility function is not dependent on endogenous beliefs, I can directly define the probabilities of a trustee playing *Share* and a trustor playing *In*, as well as their (average) first- and second-order beliefs, as follows:

$$\mathbb{P}(Share) = \int_0^1 \int_{\frac{g+\theta^I(m-1)}{2+g}}^1 f(\theta^S, \theta^I)\, d\theta^I d\theta^S = 1 - \text{marg}_{\Theta^I} F(\frac{g+\theta^I(m-1)}{2+g}), \tag{10}$$

$$\begin{aligned} \alpha_A^{Sh} &= \int_0^1 \int_{\frac{g+\theta^I(m-1)}{2+g}}^1 f_{e_A}(\theta^S, \theta^I)\, d\theta^I d\theta^S = e_A, \\ \mathbb{E}[\alpha_A^{Sh}] &= \int_0^1 e_A\, dK(e_A), \end{aligned} \tag{11}$$

$$\mathbb{P}(In) = \int_0^1 \mathbf{1}_{(\alpha^{Sh} \geq \frac{1}{2})}\, dK(e_A) = \int_{\frac{1}{2}}^1 dK(e_A), \tag{12}$$

$$\begin{aligned} \beta_B^{\varnothing} &= \mathbb{E}_{e_B}[\alpha_A^{Sh}] = \int_0^1 \alpha^{Sh}\, dK_{e_b}(e_A), \\ \mathbb{E}[\beta_B^{\varnothing}] &= \int_0^1 \beta_B^{\varnothing}\, dK(e_B). \end{aligned} \tag{13}$$

From the monotonic properties of integrals and the dependence of $\hat{\theta}^S$ on m, Proposition 2 follows trivially:

Proposition 2 *Consider two Trust Minigames with parameters m and m' and populations of inequity-averse trustees described by $F(\theta^S, \theta^I)$ and $K(e_B)$ and trustors described by $K(e_A)$. Then, for every $m' > m \geq 1$, the following relationships hold:*

- $\mathbb{P}(Share)|_{m'} \leq \mathbb{P}(Share)|_m$;
- $\mathbb{P}(In)|_{m'} \leq \mathbb{P}(Share)|_m$;
- $\mathbb{E}[\alpha_A^{Sh}]|_{m'} \leq \mathbb{E}[\alpha_A^{Sh}]|_m$;
- $\mathbb{E}[\beta_B^{\varnothing}]|_{m'} \leq \mathbb{E}[\beta_B^{\varnothing}]|_m$.

This result paints a strikingly different picture compared to the outcome described in Proposition 1 concerning Guilt Aversion. Here, as the wealth gap between the two parties widens, the likelihood of pro-social actions decreases, as do the associated beliefs. This outcome is primarily driven by the growing disadvantageous inequality for the trustee, reducing the appeal of choosing *Share* and resulting in a diminished propensity to do so. The different implications of Proposition 1 and Proposition 2 will serve as theoretical justification for the treatment effect in the experiment presented in the next section.
In the following section, I will discuss the experimental design and methodology. The experiment is a between subjects design that involves two game forms: one with $m = 1$ and another with $m = 2$. Using the results of Proposition 1 and Proposition 2, it will be possible to test the hypotheses of Guilt and Inequity Aversion. More detailed, testable predictions will be presented in Section 4. By conducting an experiment that mirrors the scenarios discussed in this chapter, we can gain valuable insights into the behaviors and preferences of participants in trust-related situations.

3 Experimental Design and Procedures

In this section I outline the details of the experimental design that allows me to test the theories of inequity aversion and guilt aversion one against the other. In this section I present the intuition behind the treatment effect, while in Section 2 I showed more formally its effect depending on which preference is assumed. The formal predictions are presented in the following section.

3.1 Experimental Design

The experiment is a between subject design and it involves a Trust Minigame in which participants assume the roles of trustor (referred to as A) and trustee (referred to as B). The main treatment is a manipulation of the first-mover payoffs, the treatment has the payoffs of player A doubled as compared to the control, while the payoffs of player B are left unchanged.

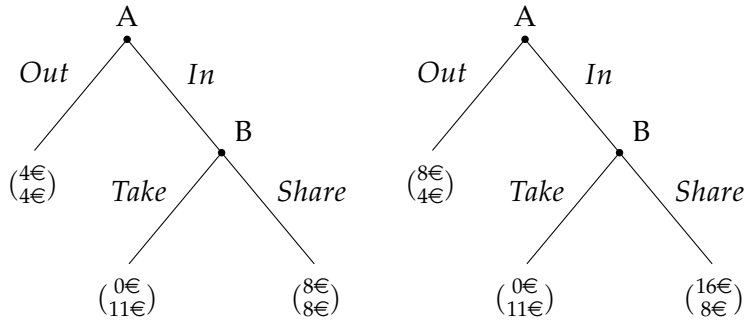


Figure 9: Game trees for the control and treated group respectively.

In this Trust Minigame player A has the option to select *Out*. If she does so, the game ends, and both participants receive 4€ each. Conversely, she can entrust player B by opting for *In*. At this point, player B has two possible responses. He can reciprocate the trust shown by selecting *Share*, resulting in both players receiving 8€. Alternatively, player B may *Take*, which yields 11€ for himself but leaves player A with 0€. Importantly, the choices of *In* or *Out* for A and *Share* or *Take* for B are made simultaneously, therefore, the game is played under the strategy method. The rules of the game for the treatment are the same, but the payoffs of player A are doubled. Figure 9 illustrates the game tree for both control and treatment.

The standard Trust Minigame involves perfect information, as A makes her choice and then B observes the choice made by A, and in the case in which A played *In*, B can then play *Share* or *Take*. Conversely, in the game employed here, the choices for A and B are made simultaneously using the strategy method. The decision to utilize the strategy method is justified, given the primary aim of this experiment: to assess the accuracy of predictions made by the two theories under examination. The theoretical justification for this method is robust, particularly given the nature of the preferences under investigation, guilt and inequity aversion, which are preferences that do not display dynamic inconsistency[5]. Moreover the strategy method has been used in other experiments invoving the Trust Minigames, including Charness and Dufwenberg (2006) and Vanberg (2008).

3.1.1 Treatment Effects

With this simple design, I can effectively compare two competing preferences, guilt and inequity aversion, one against the other. Formal testable predictions will be elaborated in the next section, building on the analysis presented in the preceding section. Here, I offer an overview of the main rationale behind my experimental design.

According to the inequity aversion model, people dislike earning more or less than others. Therefore, the inequity aversion model predicts that B experiences disutility when playing *Take*. If B opts for *Take*, he incurs a disutility is equal to $\theta_B^S(11-0)$, where θ_B^S represents his sensitivity to superiority aversion. If B's sensitivity is sufficiently high, he will prefer to play *Share*. A is aware of B's preferences, and if she believes that his sensitivity is high enough, she may choose to play *In*.[6] Therefore, the inequity aversion model effectively accounts for the occurrence of positive fractions of *In* and *Share* plays.

Similarly, the guilt aversion model can also account for these same findings. If B experiences guilt aversion, he experiences disutility when choosing *Take*. This disutility is directly related to A's disappointment. A opts for *In* if she anticipates that B is more likely to choose *Share*, then she expects to gain $\alpha_A^{Sh} \cdot 8$, where $\alpha_A^{Sh} \geq 1/2$ represents her first-order subjective belief in B selecting *Share*. In the terminal history $(In, Take)$, A's disappointment equals the difference between her expected and actual payoff, which amounts to $\alpha_A^{Sh} \cdot 8$. Consequently, the disutility experienced by B when playing *Take* is $\theta_B^G \cdot \beta_B^{Sh} \cdot 8$, where θ_B^G represents B's sensitivity to guilt, and β_B^{Sh} denotes his second-order belief—specifically, his belief about A's belief in his choice of *Share*. If both θ_B^G and β_B^{Sh} are sufficiently high, B will prefer to play *Share*, and A anticipates this and chooses *In*.

[5] See Battigalli and Dufwenberg (2009) and Battigalli et al. (2019) for more details.

[6] As assumed in the previous section, A is still considered selfish.

Given the similar predictions derive from the two models, it is challenging to discern which model drives the players' decisions in a Trust Minigame. Notably, with my design, I'm able to isolate and disentangle the effect of the two preferences as they provide opposite predictions regarding the treatment effect. The doubling of the material payoffs for A results in contrasting outcomes for the two models. Under the assumption of inequity aversion, B experiences disutility due to his inferiority aversion in the terminal histories (*Out*) and (*In*,*Share*). It's worth noting that B's inferiority aversion was absent in the control game, and his material incentives remain unchanged. In the treatment condition, *Share* becomes less appealing to B when compared to the control game. This is because his utility is now reduced by $\theta_B^I(16-8)$, where θ_B^I represents B's sensitivity to inferiority aversion. Thus, for B to choose *Share*, he needs to possess a higher θ_B^S to counterbalance the disutility arising from his inferiority aversion. This implies that only individuals with the highest sensitivity, in other words, those who are more altruistic, will opt for *Share*. Consequently, according to the inequity aversion model, the frequency of *Share* in the treatment will be lower. The same reasoning applies to the frequency of *In*, as A can anticipate B's decision-making process and behavior. Conversely, the guilt aversion model makes the opposite prediction. With A's increased payoffs, she now anticipates gaining $\alpha^{Sh} \cdot 16$ by selecting *In*. If B disappoints A, her disappointment will be more substantial, thereby proportionally increasing B's guilt as well. Consequently, there will be an additional fraction of B subjects inclined to play *Share* in the treatment. Once again, A foresees this behavior, leading to a similar increase in the choice of *In*.

The simple manipulation of the payoffs of the first-mover allows me to discern between the two models. Higher recorded frequency of *Share* in the treatment are proof that a majority of the subjects is affected by guilt when making the decision. A lower frequency is instead a sign of inequity aversion. A higher or lower frequency of *In* informs us about the beliefs of the A players about B players' preferences. First- and second-order beliefs follow a similar pattern.

3.2 Experimental procedures

For this experiment, I recruited 128 participants from the subject pool of Bocconi University through SONA.[7] The subject pool primarily consists of students from business and economics disciplines. The experiment was programmed using zTree (Fischbacher 2007) and conducted at the Bocconi Experimental Laboratory for the Social Sciences (BELSS). It was approved by Bocconi's Ethics Committee Review. Data were collected and handled in compliance with the European General Protection Regulation (GDPR, 2016/679). The average payment amounted to 11.10€, which included a 5€ show-up fee, and the experiments had an average duration of 45 minutes. Payments were disbursed through Amazon gift cards. I conducted a total of 8 sessions, with each session designated either as control or treatment. The subjects selected which session to participate in, and then the sessions were randomized to either control or treatment. To ensure an equal number of observations in both groups the session were in equal number, 4 sessions for the control and 4 sessions for the treatment. As a result, in each group, there were 32 subjects, with half of them assuming the role of A and the other half taking on the role of B. The experiment took place in February 2023.

At the outset of the experiment, each participant was randomly assigned to either role A or role B. Therefore, in each session, there were 8 players in role A and 8 players in role B, and participants

[7] Contrary to what was declared in the pre-registration, the experiment was conducted at Bocconi University instead of University of Siena. This change was due to the immediate availability of the laboratory at Bocconi, while the facility in Siena was not ready to start. Therefore, the location was changed to expedite the process.

retained their assigned roles throughout the entire experiment. In each session, participants engaged in eight rounds, with those in role A being matched with participants in role B through an absolute typed stranger matching, meaning that each participant in a given role played against every other participant in the opposite role once and only once. The repetition was done to improve the consistency of the results (Hey 2001). Once participants were situated at their computers, I read the instructions aloud and ensured that the rules were thoroughly understood. Additionally, written instructions were printed and distributed to each participant (instructions in Appendix A).

During each of the eight rounds, participants were presented with a decision and a belief elicitation task. These tasks remained the same throughout the experiment. Participants were required to make a choice for each task, and both tasks were displayed on the screen simultaneously (screenshots of the users' interface in Appendix B). In the decision task, each player had to select an action. The choice of action differed between the two roles: As had to decide between *In* and *Out*, while Bs had to choose between *Share* and *Take*. In the belief elicitation task, participants in role A were tasked with reporting their first-order beliefs, specifically, to estimate how many Bs would choose *Share*. They had nine options to choose from, ranging from $0/8$ to $8/8$. On the other hand, participants in role B were asked to report their second-order beliefs. They had to guess the estimate made by their co-player in role A, and they also had nine options to select from, ranging from $0/8$ to $8/8$. Participants' payments were determined based on the outcome of one randomly selected round they played. Additionally, they received 0.5€ for each correct estimate of guess. After the eight rounds, participants completed a non-incentivized questionnaire. This questionnaire presented eight different Trust Minigames. For each scenario, participants, in the same roles as before, were asked to indicate their choice and belief, mirroring the previous phase. Each game featured distinct parameters, including different ratios between trustor and trustee payoffs ($m = 1$, $m = 1,5$, $m = 2$ and $m = 2,5$) and gains associated with choosing *Take* ($g = 1$ and $g = \frac{1}{2}$). Finally, at the conclusion of the experiment, participants were asked to fill out a questionnaire containing demographic questions.

4 Behavioral Predictions

The primary objective of this paper is to explore the preferences of a trustee in order to gain insight into and potentially predict their behavior in real-life situations akin to a trust game. In Section 2, I conducted two separate analyses of the game, each time assuming that only one preference was relevant. This simplified analysis facilitates the analysis of each model, helps with the exposition, and yields clean results. Nevertheless, the behavioral implications derived from the analysis in Section 2 remain the primary explanation for a difference in behaviour and beliefs between the control and treated group. This is due to the fact that the results are primarily influenced by the fundamental difference in how the two preferences respond to a higher endowment of the co-player.

With my experiment, I aim to empirically test the predictions derived from the findings in Section 2.[8] These predictions naturally arise from the analysis by setting $m = 1$ for the game in the control condition and $m = 2$ for the game in the treatment condition. For each behavioral prediction, I present two hypotheses, one stemming from Guilt Aversion (a) and the other from Inequity Aversion (b). These pairs of hypotheses are always mutually exclusive.

[8] Both testable hypotheses and the experimental design have been pre-registered and can be accessed at https://doi.org/10.17605/OSF.IO/EFKQ6.

4.1 Behavioral predictions about trustee

The primary behavioral prediction under investigation pertains to the frequency at which players in the B role choose *Share*. These hypotheses represent the central testable propositions of this paper, and all other predictions stem from them.

Hypothesis 1.a. If B players are guilt-averse, then the frequency of *Share* is *higher* in the treatment than in the control.

Hypothesis 1.b. If B players are inequity-averse, then the frequency of *Share* is *lower* in the treatment than in the control.

Contrary to previous works, these predictions are not the result of an exogenous change in beliefs, but are determined by the preference over others' payoffs. This is true even for Guilt Aversion, despite being a belief-dependent preference. This is thanks to the choice of setting $m = 2$ for the treatment as pointed out at the end of Section 2.1. TThis design choice simplifies the analysis and allows me to formulate Hypothesis 1.a without the need to make assumptions about the beliefs of the B players.

4.2 Behavioral predictions about trustor

Next, I will compare the behavior of A players across treatments. According to my previous theoretical analysis, the trustor's behavior is not influenced by her preferences, as she is assumed to be selfish. Instead, her behavior is shaped by her beliefs about B's preferences and rationality. The assumption of her being selfish may seem farfetched at first glance, but given her role, it is possible to argue that her sense of guilt is not triggered, as argued in Attanasi et al. (2016). Also, the choice of the material payoffs mitigates the role of possible other-regarding preferences. For instance, if it is assumed that she is inequity-averse, it is easy to see that the utility from the terminal history $(In, Share)$ is twice the utility of (Out); this is true for every level of inequity aversion. This further supports the idea that her behavior is primarily driven by her strategic considerations instead of her preferences.

The predictions regarding the frequency of *In* in the two treatments, as suggested by the Guilt and Inequity Aversion models, are as follows:

Hypothesis 2.a. If A players believe that B players are guilt-averse, then the frequency of *In* is *higher* in the treatment than in the control.

Hypothesis 2.b. If A players believe that B players are inequity-averse, then the frequency of *In* is *lower* in the treatment than in the control.

4.3 Predictions about elicited beliefs

During the experiments, participants are asked to report their beliefs. Specifically, A players are tasked with estimating how many Bs will choose *Share*, constituting their first-order beliefs. On the other hand, B players are asked to predict A players' guesses, representing their second-order beliefs. The predictions on first- and second-order beliefs are as follows:

Hypothesis 3.a. Under the hypothesis of guilt aversion, *A* players' average reported first-order are *higher* in the treatment than in the control.

Hypothesis 3.b. Under the hypothesis of inequity aversion, *A* players' average reported first-order beliefs are *lower* in the treatment than in the control.

Hypothesis 4.b. Under the hypothesis of guilt aversion, *B* players' average reported second-order beliefs are *higher* in the treatment than in the control.

Hypothesis 4.b. Under the hypothesis of inequity aversion, *B* players' average reported second-order beliefs are *lower* in the treatment than in the control.

These hypotheses require a degree of sophistication from both *A* and *B* players. Regarding hypotheses 3.a and 3.b, *A* players have to believe that *B* players are guilt- or inequity-averse, respectively. For hypotheses 4.a and 4.b, an additional step is required. In order to change their beliefs, *B* players have to recognize that *A* players believe them to be guilt- or inequity-averse, respectively.

5 Results

This section presents the results obtained in my experiment.[9] In Section 5.1 I present the descriptive statistics, showing a strong treatment effect on both behavior and beliefs for participants in both roles. The change in behavior and beliefs is in line with the prediction derived from the inequity aversion model, and therefore opposite to what predicted by the guilt aversion model. In Section 5.2 I analyse the data using panel regression with random effects. The results of the regression confirm the strong treatment effect.

5.1 Descriptive and Preliminary Evidences

This analysis begins by reporting the frequency of the actions played and the average beliefs reported by the participants. Starting from the behaviour of the trustee, the main focus of this work, in the control group, the action *Share* was played 44,1% of the time (113/256), while in the treated group it was played only 20,3% of the time (52/256). A one-sided t-test and a non parametric Mann-Whitney test performed by averaging the eight choices of each subject show that the difference is statistically significant at 1%, as reported in Table 1.[10] The frequency of *Share* is halved from the control to the treatment, which means subjects behave more selfishly when there is inequality that is not in their favor, despite the increase in the potential loss for the trustor. This first result shows a strong treatment effect, which is in line with the predictions given by inequity aversion. This is the most important result of this paper because it shows the direct effect on behavior of jointly increasing inequality and guilt.

[9] The statistical analysis in this section follows the pre-analysis plan available at https://osf.io/efkq6. Further robustness tests and additional figures are presented in Appendices C and D

[10] In Table 1, I report the average over the eight rounds of behavior and beliefs for each individual, since I cannot use every data point because they are not independent. In Table D.2 in Appendix D, I perform the same analysis separately for every round. The results remain largely consistent with those presented in Table 1, being significant at 5% starting from the third round.

Regarding the behaviour of trustors, the action *In* was played 34,4% of the times (88/256) in the control group, while it was played 14,1% of the times (36/256) in the treated group. Also, this difference is statistically significant at 1%, as reported in Table 1. Once again the frequency of the pro-social actions is halved going from the control to the treatment. This signals a strong treatment effect in favor of the hypothesis supported by inequity aversion, showing that the subjects believe that *B*s behave as inequity-averse, a fact that is indeed confirmed by the previous result.[11] The frequencies of the actions played by both players *A* and *B* are shown in Figure 10.

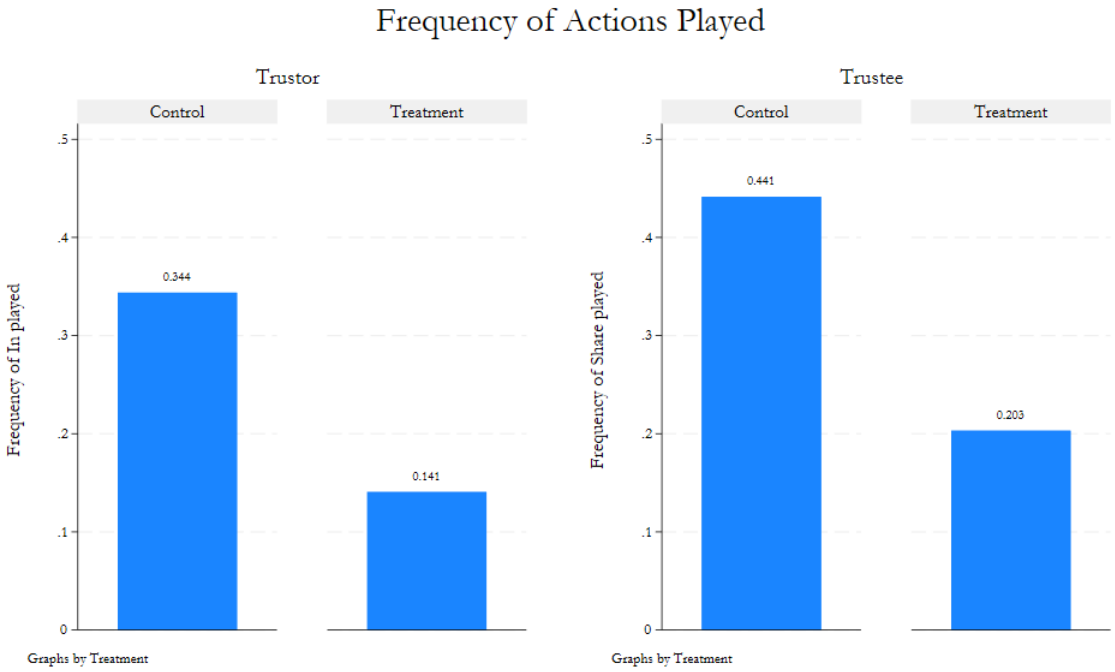


Figure 10: Frequency of *In* and *Share* respectively, in the two treatments.

A similar trend is displayed by first- and second-order beliefs, as shown in Figure 11. The lack of trust is also reflected in trustors' reported first-order beliefs. The average reported first-order belief is 0,30 in the control group and 0,20 in the treated group. This difference is significant at 1%. Again, this result is in favor of the inequity aversion model, and it's consistent with the hypothesis that players *A* believe that players *B* are inequity-averse and therefore expect a drop in the frequency of *Share* in the treatment. First-order beliefs and behavior are strongly correlated (Spearman's correlation coefficient $r_s = 0,55$, $p = 0,000$), as is expected from rational agents. A similar depiction comes from trustees' reported second-order beliefs, with an average of 0,35 and 0,20 in the control and treated groups, respectively, difference significant at 5%. Also, this result confirms the hypothesis in favor of inequity aversion, with the *B* players correctly identifying the change in *A* players' beliefs.

An additional result is the strong and positive correlation between second-order beliefs and behavior (Spearman's correlation coefficient, $r_s = 0,45$, $p = 0,000$). Inequity aversion, in its formu-

[11] Difference in behavior between treatment conditions is stable between sessions for both trustees and trustors as shown in Figure D.3 in Appendix D.

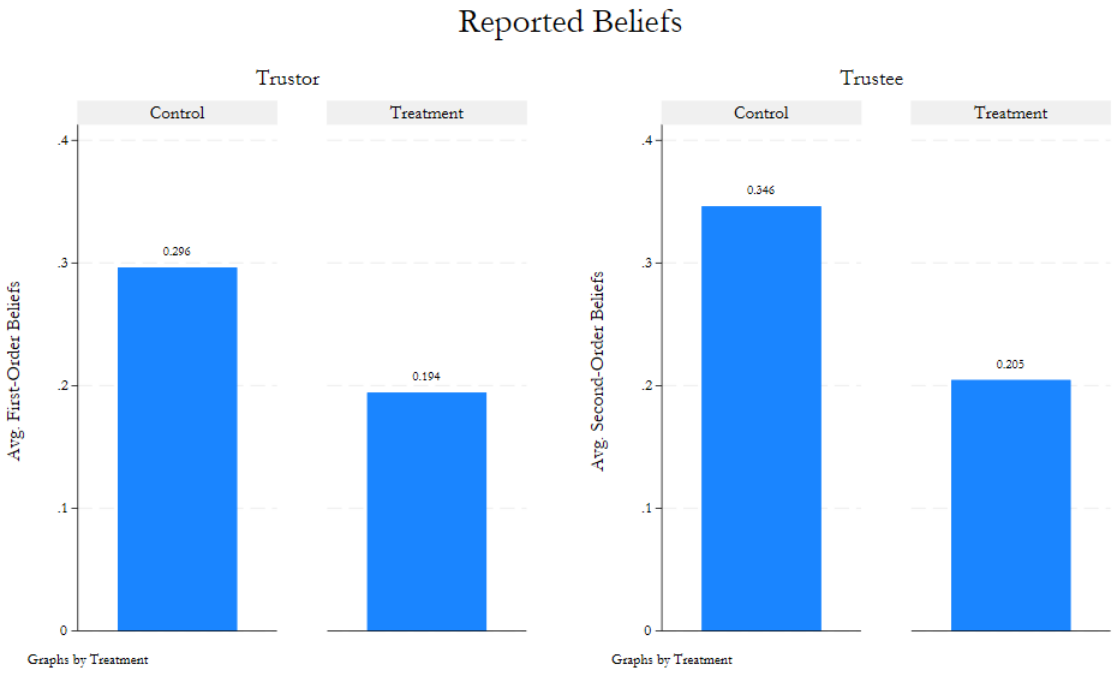


Figure 11: Average first- and second-order beliefs, respectively, in the two treatments.

Table 1: One-Sided T-tests and Mann-Whitney test across the two groups.

	Control	Treatment	Diff.	p-value	MW p-value	obs.
Freq. *In*	0.34	0.14	0.20***	0.003	0.018	64
Freq. *Share*	0.44	0.20	0.24***	0.003	0.006	64
First-Order Belief	0.30	0.20	0.10**	0.035	0.030	64
Second-Order Belief	0.35	0.20	0.14***	0.006	0.006	64

Note: *** p<0.01, ** p<0.05, * p<0.1

lation, cannot predict this correlation, while guilt aversion can. The positive correlation between second-order beliefs and behavior has been reported and used as proof in favor of guilt aversion in many other works, first among those Charness and Dufwenberg (2006). In light of the results of this experiment, despite this evidence in favor of guilt aversion, it appears that inequity aversion is the main behavioural driver and this is correctly recognized by the majority of the participants. It is also important to point out that my experimental design is not best suited for studying this correlation since the variation in second-order beliefs is endogenous. I defer to the existing literature for an in-depth analysis of the exogenous variation of beliefs in a Trust Minigame.

5.2 Regression Analysis

The preliminary investigation of the data reveals that all four behavioral hypotheses supporting inequity aversion are indeed confirmed, while those favoring guilt aversion, being mutually exclusive, must be rejected. This suggests that preference for equity prevails in the context of the experiment where there is difference in the stakes of the two subjects involved. Furthermore, the treatment effect remains significant even after conducting a linear panel regression with random effects at the individual level. As illustrated in Table 2, the treatment effect is consistently negative and significant across all dependent variables of interest. In the initial two columns of Table 2, the dependent variables represent the choices of trustors and trustees, respectively. For both the frequency of *In* and *Share*, the treatment reduces the probability of these choices being made by 22 and 21 percentage points from a baseline of 34% and 44%, respectively. In the last two columns, beliefs serve as the dependent variables, encompassing both first-order and second-order beliefs. The treatment results in a reduction of 11 and 13 percentage points for first-order and second-order beliefs, respectively. These findings underscore the strength of the treatment effect. I run further robustness tests in Appendix D. There I perform a cross-sectional regression, averaging the choice of each subject, and in another panel regression I include the lagged partner's choice to account for the feedback received between rounds. The consistency of the treatment effect across different statistical analyses underscores its robustness and reliability. These results contribute not only to our understanding of behavioral dynamics in the context of inequity and trust but also offer valuable insights for policymakers and practitioners seeking to design interventions that promote fairness and cooperation in social and economic interactions.

Table 2: Panel regressions (GLS) for choices and beliefs.

	Choices		Beliefs	
	(Role A)	(Role B)	(First-order)	(Second-order)
Treatment	-0.222***	-0.213**	-0.105***	-0.126**
	(0.033)	(0.085)	(0.038)	(0.053)
Round	-0.010**	-0.001	-0.005	-0.017***
	(0.004)	(0.008)	(0.008)	(0.006)
Gender	0.015	0.112	0.128***	0.075
	(0.046)	(0.110)	(0.042)	(0.067)
Observations	512	512	512	512
Demographics	YES	YES	YES	YES

Note: Standard deviations/errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

6 Conclusions

This paper presents a theory-driven experiment to test the model of guilt aversion against the model of inequity aversion in the context of a Trust Minigame. The contribution is both theoretical and empirical, extending the discussion initiated by Vanberg's response to Charness and Dufwenberg (2006). It offers a clear method to test the two models and provides strong evidence

in favor of inequity aversion.

The first contribution of this paper is theoretical. I showed that the two models under investigation react differently to an increasing inequality between trustor and trustee. As the difference between the initial allocations of the two parties increases, inequity aversion predicts a decreasing frequency of pro-social behavior, while guilt aversion predicts the opposite. This divergence in predictions is rooted in the intrinsic characteristics of the two preferences and how they respond to an increase in the co-player's payoff. Inequity aversion amplifies the sense of *envy* with rising inequality, leading to a reduction in pro-social behavior from the trustee. In contrast, guilt aversion heightens the magnitude of the trustor's potential disappointment, resulting in an increase in pro-social behavior as the trustee wants to avoid a greater sense of guilt. Understanding these distinct reactions sheds light on the underlying mechanisms driving behavior in trust-related situations and has allowed me to design an experiment capable of testing the predictions provided by the two preferences.

The data gathered from the experiment reveal a significant decrease in pro-social behavior in the treated group, as well as shifts in beliefs about pro-social behavior for both trustors and trustees. The frequency of *Share* played drops from 44.1% in the control group to 20.3% in the treated group. Similarly, the frequency of *In* played decreases from 34.4% in the control to 14.1% in the treatment. The frequency of pro-social behavior is halved when the initial allocation of the trustor is doubled. Both first- and second-order beliefs consistently exhibit a similar reduction. The results align with the predictions derived from inequity aversion in the theoretical analysis, offering robust support in favor of inequity aversion. Particularly, they underscore the significant role of inferiority aversion in people's decision-making in situations affected by inequality between the two parties.

References

Arrow, K. J. (1972). Gifts and exchanges. Philosophy & Public Affairs, 343–362.

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

Battigalli, P., R. Corrao, and M. Dufwenberg (2019). Incorporating belief-dependent motivation in games. Journal of Economic Behavior & Organization 167, 185–218.

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

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

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

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

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. Games and economic behavior 10(1), 122–142.

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

Ciriolo, E. (2007). Inequity aversion and trustees' reciprocity in the trust game. European Journal of Political Economy 23(4), 1007–1024.

Della Lena, S., E. Manzoni, and F. Panebianco (2023). On the transmission of guilt aversion and the evolution of trust. Games and Economic Behavior 142, 765–793.

Di Bartolomeo, G., M. Dufwenberg, and S. Papa (2023). Promises and partner-switch. Journal of the Economic Science Association, 1–13.

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

Dirks, K. T. and D. L. Ferrin (2002). Trust in leadership: meta-analytic findings and implications for research and practice. Journal of applied psychology 87(4), 611.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. Journal of Economic Behavior & Organization 48(1), 57–69.

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

Ellingsen, T. and M. Johannesson (2004). Promises, threats and fairness. The Economic Journal 114(495), 397–420.

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

Fehr, B. (1988). Prototype analysis of the concepts of love and commitment. Journal of personality and social psychology 55(4), 557.

Fehr, E. and G. Charness (2023). Social preferences: fundamental characteristics and economic consequences.

Fehr, E. and K. M. Schmidt (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, 171–178.

Fukuyama, F. (1996). Trust: The social virtues and the creation of prosperity. Simon and Schuster.

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

Hey, J. D. (2001). Does repetition improve consistency? Experimental economics 4, 5–54.

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

Knack, S. and P. Keefer (1997). Does social capital have an economic payoff? a cross-country investigation. The Quarterly journal of economics 112(4), 1251–1288.

Kreps, D. M. et al. (1990). Corporate culture and economic theory. Perspectives on positive political economy 90(109-110), 8.

Rothstein, B. and E. M. Uslaner (2005). All for all: Equality, corruption, and social trust. World politics 58(1), 41–72.

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

A Instructions

The instructions are translated from Italian.

A.1 Instruction: Main Task (Control)

Welcome to BELSS (Bocconi Experimental Laboratory for the Social Sciences). Thank you for your participation in this experiment. Feel free to ask us questions by raising your hand. Please refrain from talking to other participants during the experiment.

The experiment consists of 3 phases: one decision-making phase and 2 questionnaires. At the beginning of the experiment, you will be randomly assigned to either role A or role B. If you have been assigned to role A, you will be paired with a participant in role B, and vice versa. You will remain in the same role for the entire duration of the session.

Decision-making Phase

The decision-making phase consists of **8 independent rounds**. In each round, you will be paired with a different participant. **You will never interact twice with the same person**. No participant will know the identity of the individuals they interacted with during the experiment or the choices made during previous rounds.

During the decision-making phase, the participant in role A chooses between CONTINUE and STOP. Simultaneously, the participant in role B chooses between SHARE and TAKE. If A chooses STOP, both A and B receive €4 each. If A chooses CONTINUE and B chooses SHARE, both receive €8. If A chooses CONTINUE and B chooses TAKE, B receives €11 while A receives €0. These options are summarized in the table below:

	A earns	**B earns**
If A chooses STOP	€4	€4
If A chooses CONTINUE and B SHARE	€8	€8
If A chooses CONTINUE and B TAKE	€0	€11

Estimates and Conjectures

Before making your choice, we will ask you to make an estimate (if you are in role A) or a speculation (if you are in role B). In detail:

- We will ask A to estimate the number of B participants who will choose SHARE in the current round. Remember that there are 8 participants in that role, so the estimate ranges from 0/8 to 8/8. The accuracy of A's estimate depends on the choices made by all B participants during the current round, and a correct estimate results in a gain of €0.50.
- We will ask B to guess the estimate of the number of B participants who have chosen SHARE made by the participant A with whom they are paired. This means that the conjecture ranges from 0/8 to 8/8. B's accuracy depends only on the estimate made by the participant A they are paired with during that round, and a correct conjecture results in a gain of €0.50.

Payments

At the end of the experiment the computer will randomly select one of the 8 rounds. You will be paid according to the choices made by you and your partner during that round. The amounts are those summarized in the table. Moreover, you will be paid 0.5€ for each correct estimate/conjecture during all 8 rounds. In addition, you will receive €5 for participating in this session and answering the final questionnaires. At the end of the session, you will be paid individually and privately.

A.2 Instruction: Main Task (Treatment)

As in the control with the following exception:

During the decision-making phase, the participant in role A chooses between CONTINUE and STOP. Simultaneously, the participant in role B chooses between SHARE and TAKE. If A chooses STOP, A receives €8 and B receives €4. If A chooses CONTINUE and B chooses SHARE, A receives €16, while B receives €8. If A chooses CONTINUE and B chooses TAKE, B receives €11, while A receives €0. These options are summarized in the table below:

	A earns	**B earns**
If A chooses STOP	€8	€4
If A chooses CONTINUE and B SHARE	€16	€8
If A chooses CONTINUE and B TAKE	€0	€11

A.3 Instructions: Hypothetical Scenarios (Role A [Role B])

Now we will present you with some hypothetical scenarios, like those previously encountered, and we will ask you how you would behave. You will not receive any payment for the value entered, nor will these values bring any monetary transfer to any of the participants in role B [A], but we ask you to give as much consideration as you did previously.

Like you did previously, you will have to choose between CONTINUE and STOP [SHARE and TAKE] and make an estimate about B's choices [a guess about A's estimate]. These scenarios have different remunerations, which are summarized in tables.

A.4 List of the Hypothetical Scenarios

- $m = 1$ and $g = 1$:

	A earns	**B earns**
If A chooses STOP	€4	€4
If A chooses CONTINUE and B SHARE	€8	€8
If A chooses CONTINUE and B TAKE	€0	€12

- $m = 1,5$ and $g = 1$:

	A earns	**B earns**
If A chooses STOP	€6	€4
If A chooses CONTINUE and B SHARE	€12	€8
If A chooses CONTINUE and B TAKE	€0	€12

- $m = 2$ and $g = 1$:

	A earns	**B earns**
If A chooses STOP	€8	€4
If A chooses CONTINUE and B SHARE	€16	€8
If A chooses CONTINUE and B TAKE	€0	€12

- $m = 2,5$ and $g = 1$

	A earns	**B earns**
If A chooses STOP	€10	€4
If A chooses CONTINUE and B SHARE	€20	€8
If A chooses CONTINUE and B TAKE	€0	€12

- $m = 1$ and $g = 1/2$:

	A earns	**B earns**
If A chooses STOP	€4	€4
If A chooses CONTINUE and B SHARE	€8	€8
If A chooses CONTINUE and B TAKE	€0	€10

- $m = 1,5$ and $g = 1/2$:

	A earns	B earns
If A chooses STOP	€6	€4
If A chooses CONTINUE and B SHARE	€12	€8
If A chooses CONTINUE and B TAKE	€0	€10

- $m = 2$ and $g = 1/2$:

	A earns	B earns
If A chooses STOP	€8	€4
If A chooses CONTINUE and B SHARE	€16	€8
If A chooses CONTINUE and B TAKE	€0	€10

- $m = 2,5$ and $g = 1/2$

	A earns	B earns
If A chooses STOP	€10	€4
If A chooses CONTINUE and B SHARE	€20	€8
If A chooses CONTINUE and B TAKE	€0	€10

B Subjects' Computer Interface

The following figures present screenshots of the main user interface. Figures D.1 and D.2 report the decision interface faced by active players in the decision-making stage for the players in role A and B respectively.

C Additional Figures and Tables

In this section, I present additional figures and a balance table comparing the control and treated groups. In particular, Table D.1 shows the descriptive statistics of both samples, which result in being mostly balanced, with the exception of gender for the trustors. The variable Gender is significantly correlated with first-order beliefs but not significantly correlated with any other variable.

Table D.1 reports the behavioral traits of the participants. These traits were determined based on their choices in the hypothetical scenarios presented after the decision experiment. I categorized participants as *Behave as Guilt Averse* if their behavior was consistent with theoretical predictions derived assuming guilt aversion. In practice, if a participant switched behavior only once from *Out/Take* to *In/Share* as m increased, I characterized them as *Behave as Guilt Averse*. Notice that this characterization has different implications for the trustor and the trustee, but I use the same term for both for the sake of parsimony. Using the same logic, I characterized participants as *Behave as Inequity Averse* if their behavior was consistent with theoretical predictions derived assuming inequity aversion. The characterizations *Always play Out/Take* and *Always play In/Share* should be self-explanatory. As we can see from Table D.1, the behavioral types are fairly balanced across treatments, with the only exception being the trustee who always played *Out* or *Take*. Another interesting observation is that there are more participants categorized as *Behave as Inequity Averse* than participants categorized as *Behave as Guilt Averse*. This is in line with the main result of this paper.

D Robustness Tests

D.1 Choices and Beliefs by Session

In Figure D.3 I report frequencies of the action played by both players A and B for each session. This graph shows that the frequency of *In* and *Share* in each session of the control was higher than in any session of the treatment. This further underscores the results, demonstrating that they are not driven by outlier behavior in any particular session. The division between the control and treatment sessions is less pronounced when we look at the beliefs, especially for the first-order beliefs, as shown in Figure D.4.

D.2 Choices and Beliefs by Round

Here, I display the choices made (Figures D.5 and D.6) and beliefs reported (Figures D.7 and D.8) round by round. The results are consistent with the main findings on a round-by-round basis. It is also noticeable from Table D.2 that the behavior and beliefs of the participants stabilize quickly. The differences between treatments are already statistically significant starting from round 3.

D.3 Cross-sectional Regression

As an additional robustness test for the panel regression presented in Table 2, I conducted a cross-sectional regression, averaging the choices and beliefs of each round for each individual. The results are presented in Table D.3. The variable *Treatment* is negative and statistically significant for all of the dependent variables of interest, confirming the main result of this work.

D.4 Panel regression controlling for feedback between rounds

During the experiment the participants received feedback about the outcome after each round. This could cause some issue, since the decisions in a given round could be influenced by the outcome of the previous round. In order to control for this, I include the choice made by the participant's partner in the previous round as a covariate. After including the lagged variable we observe that the coefficients of the treatment are still negative and statistically significant at 1-5%. This is reassuring and strengthen the result presented in Section 5. As anticipated, past partner's choices are very relevant. The lagged variable is significant at 1% for every dependent variable, with the only exception being the choice of B players. Also this evidence is in line with the predictions given by inequity aversion. First, the choices of B players don't depend of past choices, signaling the absence of belief dependent preferences like guilt or reciprocity. The positive relationship between the lagged variable and the choice of A players show that their behavior largely depends on strategic considerations, A players are more likely to trust if their trust was repaid in the previous round. The lagged variable is also positively correlated with beliefs, both first- and second-order. This show that participants correctly update their beliefs based on their partners' choices.

Table D.1: Descriptive statistics and Balance by Treatment

	Trustor			Trustee		
	(1)	(2)	(3)	(4)	(5)	(6)
	Control	Treatment	Diff. (1)-(2)	Control	Treatment	Diff. (3)-(4)
Demographics						
Gender	0.59	0.31	0.28**	0.50	0.41	0.09
	(0.50)	(0.47)	(0.12)	(0.51)	(0.50)	(0.13)
Age	20.72	20.84	-0.12	20.97	20.75	0.22
	(1.69)	(1.83)	(0.44)	(1.99)	(2.02)	(0.50)
Year of Study	2.44	2.56	-0.12	2.44	2.66	-0.22
	(1.34)	(1.39)	(0.34)	(1.52)	(1.54)	(0.38)
Lab Exp	1.75	2.31	-0.56	2.00	2.03	-0.03
	(1.90)	(2.39)	(0.54)	(3.42)	(3.69)	(0.89)
Studies						
Econ & Fin	0.31	0.59	-0.28**	0.50	0.41	0.09
	(0.47)	(0.50)	(0.12)	(0.51)	(0.50)	(0.13)
Management	0.44	0.28	0.16	0.34	0.47	-0.12
	(0.50)	(0.46)	(0.12)	(0.48)	(0.51)	(0.12)
Law	0.03	0.12	-0.09	0.03	0.06	-0.03
	(0.18)	(0.34)	(0.07)	(0.18)	(0.25)	(0.05)
Statistics	0.06	0.00	0.06	0.00	0.00	0.00
	(0.25)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)
Inter. Relationships	0.03	0.00	0.03	0.06	0.00	0.06
	(0.18)	(0.00)	(0.03)	(0.25)	(0.00)	(0.04)
Behavioural Traits						
Always play Out/Take	0.38	0.38	0.00	0.28	0.50	-0.22*
	(0.49)	(0.49)	(0.12)	(0.46)	(0.51)	(0.12)
Always play In/Share	0.00	0.00	0.00	0.12	0.09	0.03
	(0.00)	(0.00)	(0.00)	(0.34)	(0.30)	(0.08)
Behave as Guilt Averse	0.09	0.09	0.00	0.00	0.03	-0.03
	(0.30)	(0.30)	(0.07)	(0.00)	(0.18)	(0.03)
Behave as Inequity Averse	0.19	0.34	-0.16	0.16	0.12	0.03
	(0.40)	(0.48)	(0.11)	(0.37)	(0.34)	(0.09)
N	32	32	64	32	32	64

Note: Standard deviations/errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table D.2: One-Sided T-tests across the two groups by round.

Round 1	Control	Treatment	Diff.	p-value
Freq. *In*	0.34	0.19	0.16*	0.08
Freq. *Share*	0.44	0.16	0.28**	0.01
First-Order Belief	2.12	2.47	-0.34	0.72
Second-Order Belief	2.97	2.34	0.62	0.14
Round 2	Control	Treatment	Diff.	p-value
Freq. *In*	0.31	0.19	0.12	0.13
Freq. *Share*	0.41	0.25	0.16*	0.09
First-Order Belief	2.12	1.84	0.28	0.30
Second-Order Belief	2.78	1.97	0.81*	0.08
Round 3	Control	Treatment	Diff.	p-value
Freq. *In*	0.41	0.19	0.22**	0.03
Freq. *Share*	0.47	0.22	0.25**	0.02
First-Order Belief	2.22	1.34	0.88*	0.06
Second-Order Belief	3.03	1.97	1.06**	0.04
Round 4	Control	Treatment	Diff.	p-value
Freq. *In*	0.38	0.12	0.25**	0.01
Freq. *Share*	0.47	0.16	0.31***	0.00
First-Order Belief	2.50	1.53	0.97**	0.03
Second-Order Belief	3.38	1.47	1.91***	0.00

Round 5	Control	Treatment	Diff.	p-value
Freq. *In*	0.34	0.12	0.22**	0.02
Freq. *Share*	0.47	0.22	0.25**	0.02
First-Order Belief	2.50	1.38	1.12**	0.02
Second-Order Belief	2.88	1.41	1.47***	0.00
Round 6	Control	Treatment	Diff.	p-value
Freq. *In*	0.34	0.03	0.31***	0.00
Freq. *Share*	0.44	0.25	0.19*	0.06
First-Order Belief	2.62	1.22	1.41**	0.01
Second-Order Belief	2.78	1.34	1.44***	0.00
Round 7	Control	Treatment	Diff.	p-value
Freq. *In*	0.31	0.22	0.09	0.20
Freq. *Share*	0.50	0.19	0.31***	0.00
First-Order Belief	2.62	1.41	1.22**	0.01
Second-Order Belief	2.16	1.44	0.72*	0.09
Round 8	Control	Treatment	Diff.	p-value
Freq. *In*	0.31	0.06	0.25***	0.00
Freq. *Share*	0.34	0.19	0.16*	0.08
First-Order Belief	2.25	1.25	1.00**	0.01
Second-Order Belief	2.19	1.16	1.03**	0.03

Table D.3: Linear regression (cross-sectional) for choices and beliefs averaged by subject.

	(A Choice)	(B Choice)	(FO Belief)	(SO Belief)
Treatment	-0.222**	-0.213**	-0.843*	-1.005**
	(0.088)	(0.096)	(0.476)	(0.470)
Gender	0.015	0.112	1.024*	0.603
	(0.083)	(0.115)	(0.521)	(0.596)
Age	0.010	0.084*	0.141	0.194
	(0.034)	(0.047)	(0.158)	(0.166)
Observations	64	64	64	64
Demographics	YES	YES	YES	YES

Note: Standard deviations/errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table D.4: Panel regressions (GLS) for choices and beliefs including past partner's choice.

	Choices				Beliefs			
	(Role A)	(Role A	(Role B)	(Role B)	(First-order)	(First-orde)	(Second-order)	(Second-order)
main								
Treatment	-0.163***	-0.185***	-0.229***	-2.004**	-0.098*	-0.100***	-0.142***	-0.127**
	(0.060)	(0.044)	(0.065)	(0.843)	(0.059)	(0.033)	(0.050)	(0.053)
Lag Partner' Choice	0.188***	0.186***	0.014	0.132	0.099***	0.100***	0.046***	0.044***
	(0.049)	(0.049)	(0.024)	(0.248)	(0.018)	(0.018)	(0.010)	(0.010)
Round	-0.012**	-0.012**	-0.005	-0.057	-0.002	-0.002	-0.017***	-0.017***
	(0.006)	(0.006)	(0.009)	(0.095)	(0.006)	(0.006)	(0.006)	(0.006)
Gender		0.005		1.526		0.120***		0.082
		(0.044)		(1.166)		(0.045)		(0.071)
Age		0.005		0.559		0.010		0.021**
		(0.024)		(0.397)		(0.016)		(0.009)
Year of Study		0.012		-0.368		-0.015		-0.035**
		(0.047)		(0.522)		(0.022)		(0.017)
Lab Exp		0.018		-0.051		0.008		0.017***
		(0.015)		(0.100)		(0.014)		(0.006)
Econ and Fin		0.109		-0.813		0.098		0.115**
		(0.185)		(3.033)		(0.130)		(0.047)
Management		0.055		-1.060		0.097		0.213**
		(0.206)		(3.396)		(0.145)		(0.086)
Law		0.188		-3.163		0.277		-0.091
		(0.226)		(2.513)		(0.173)		(0.067)
Statistics		0.420**		0.000		0.175		0.000
		(0.191)		(.)		(0.146)		(.)
Inter. Relationships		0.077		0.000		-0.137		0.137
		(0.202)		(.)		(0.167)		(0.126)
Constant	0.320***	0.059	0.465***	-10.659	0.266***	-0.070	0.412***	-0.144
	(0.073)	(0.491)	(0.060)	(10.187)	(0.046)	(0.373)	(0.065)	(0.258)
Observations	448	448	448	434	448	448	448	448
Demographics	NO	YES	NO	YES	NO	YES	NO	YES

Note: Standard deviations/errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Figure D.1: Decision Interface, Role A

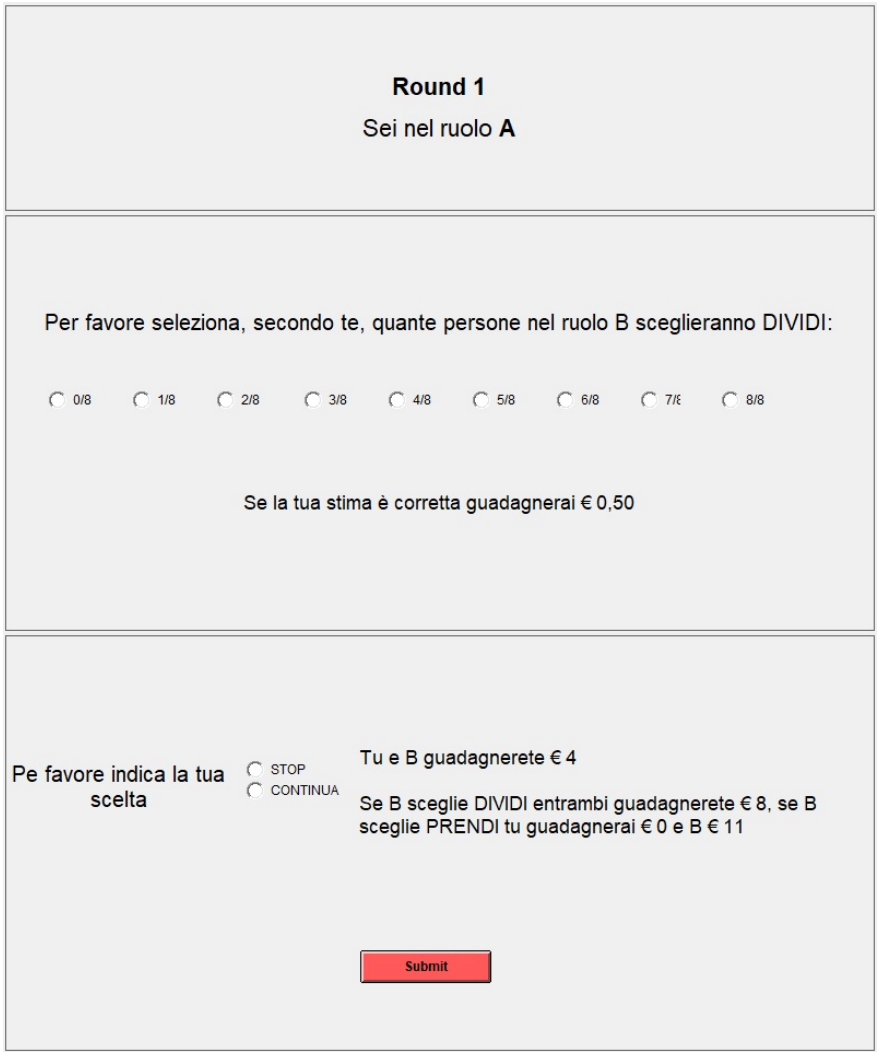
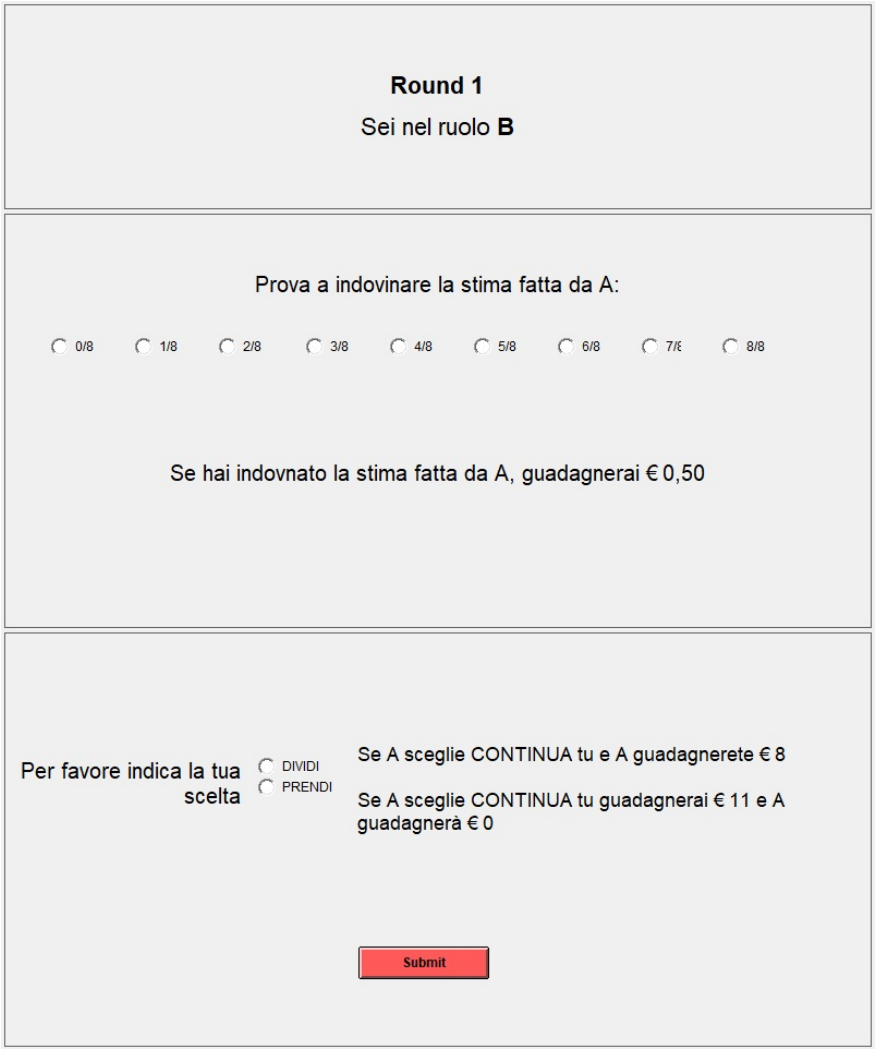


Figure D.2: Decision Interface, Role B



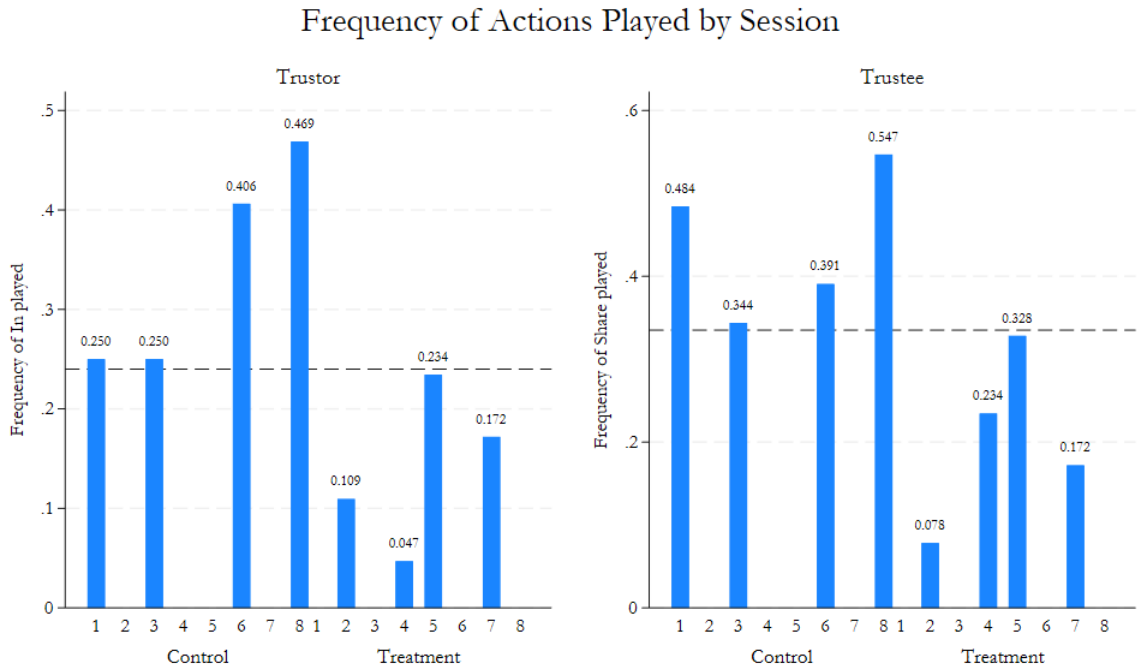


Figure D.3: Frequency of *In* and *Share* respectively, in each session divided by treatment.

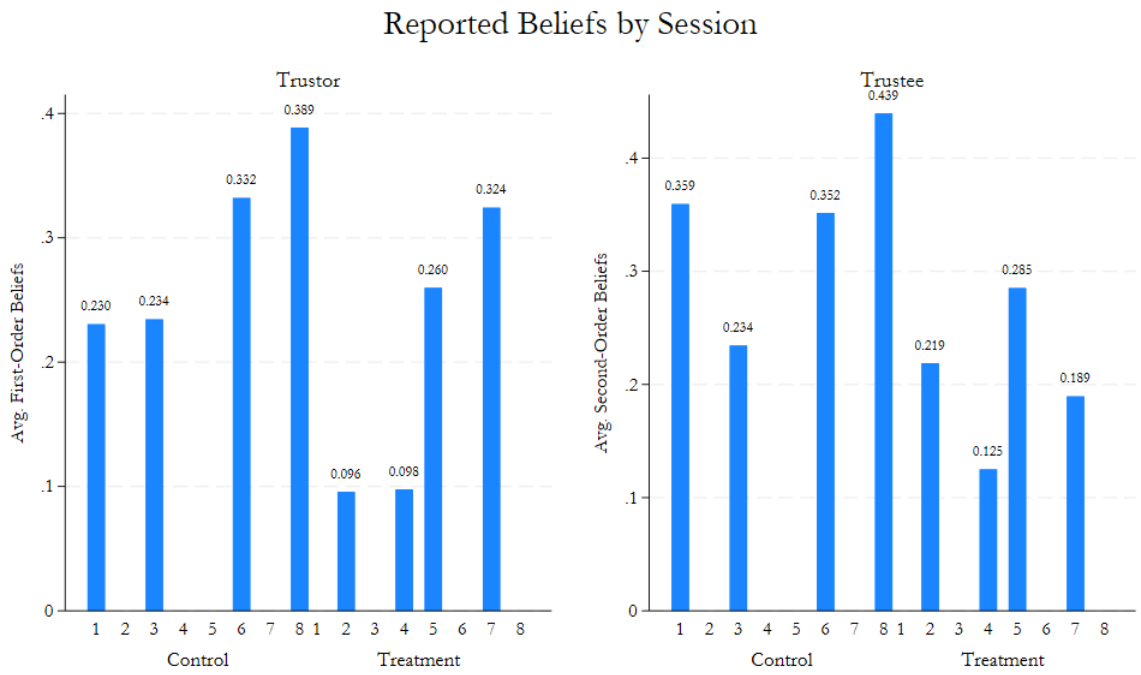


Figure D.4: Average first- and second-order beliefs, respectively, in each session divided by treatment.

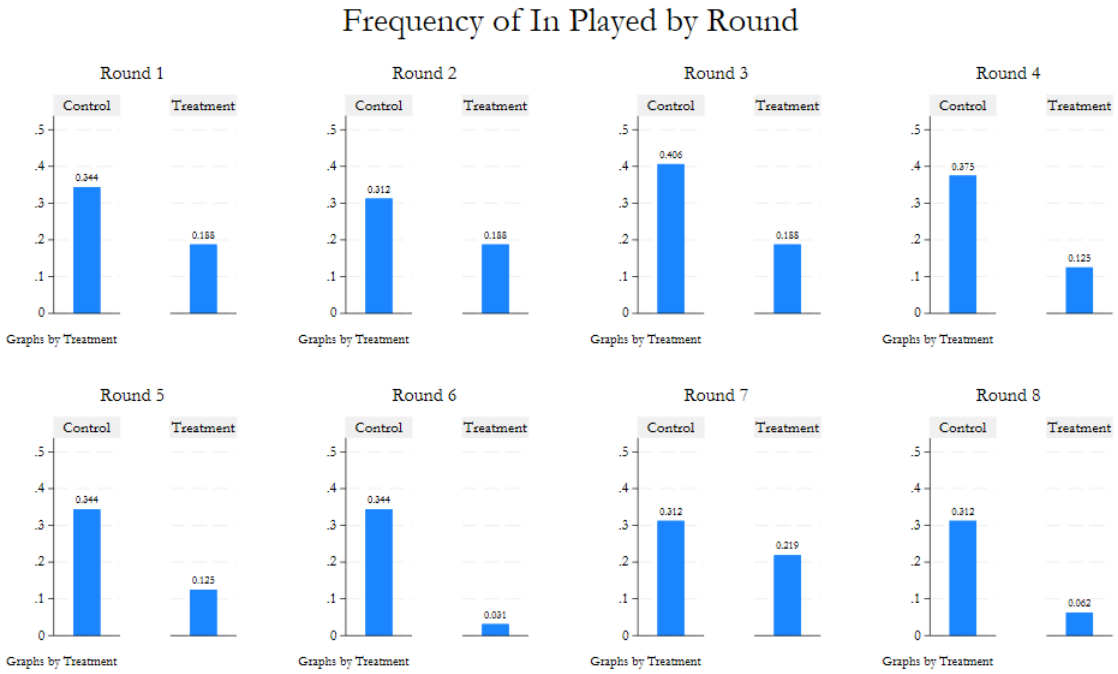


Figure D.5: Frequency of *In* in the two treatments, round-by-round.

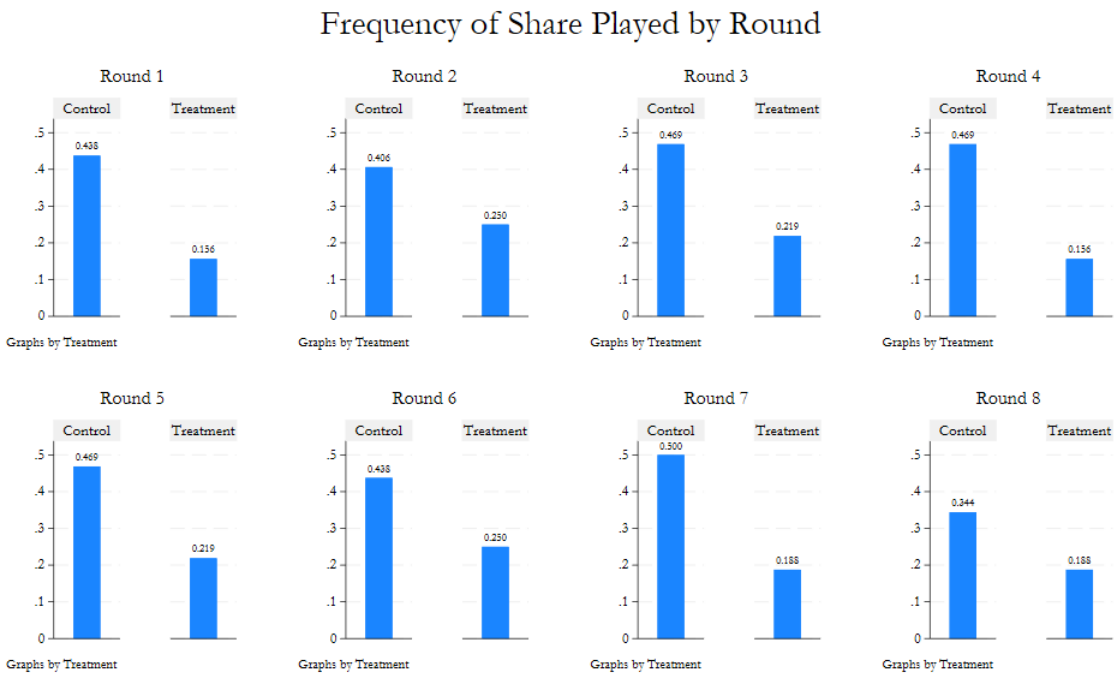


Figure D.6: Frequency of *Share* in the two treatments, round-by-round.

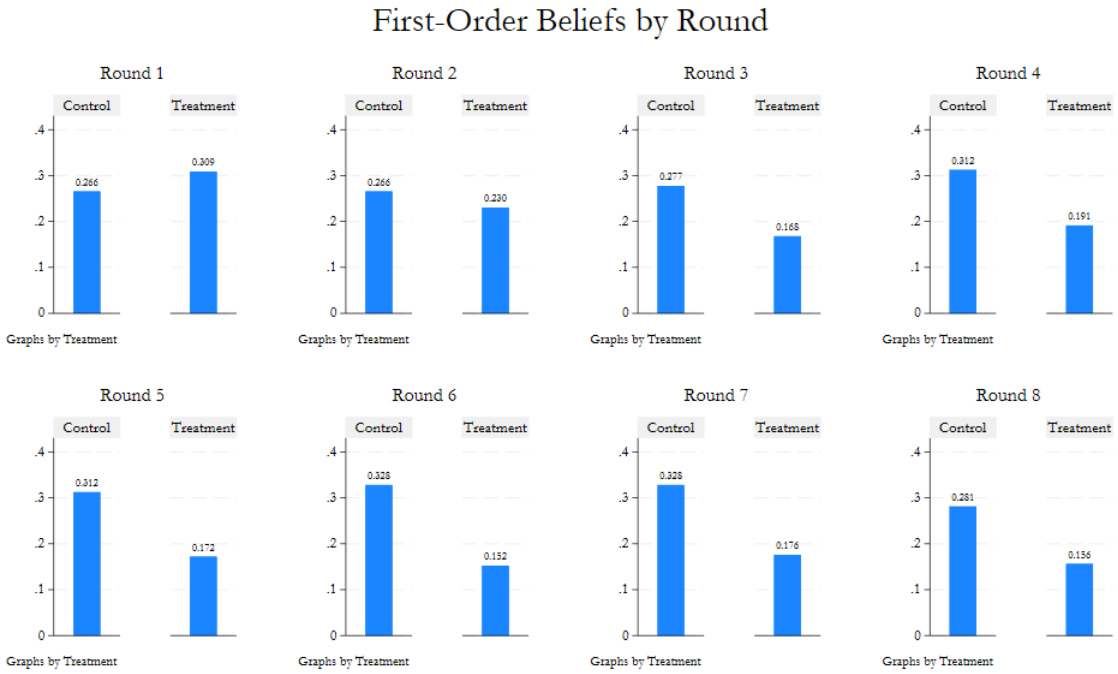


Figure D.7: Reported First-Order Beliefs in the two treatments, round-by-round.

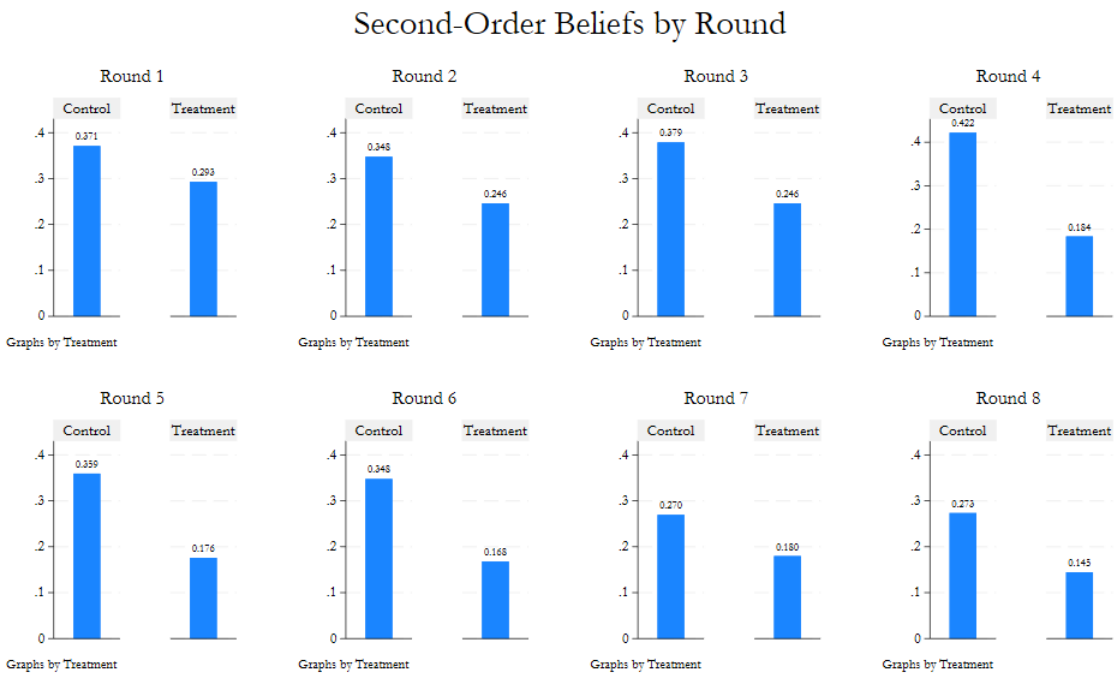


Figure D.8: Reported Second-Order Beliefs in the two treatments, round-by-round.