DEPARTMENT OF ECONOMICS

UNIVERSITY OF MILAN - BICOCCA

WORKING PAPER SERIES

Incomplete Information Models of Guilt Aversion in the Trust Game

Giuseppe Attanasi, Pierpaolo Battigalli, Elena Manzoni

WP No. 246 – June 2013

Dipartimento di Economia Politica
Università degli Studi di Milano - Bicocca
http://dipeco.economia.unimib.it

Incomplete Information Models of Guilt Aversion in the Trust Game*

Giuseppe Attanasi
University of Strasbourg, BETA

Pierpaolo Battigalli
Bocconi University, IGIER

Elena Manzoni
University of Milan-Bicocca

Abstract

In the theory of psychological games it is assumed that players' preferences on material consequences depend on endogenous beliefs. Most of the applications of this theoretical framework assume that the psychological utility functions representing such preferences are common knowledge. But this is often unrealistic. In particular, it cannot be true in experimental games where players are subjects drawn at random from a population. Therefore an incomplete-information methodology is called for. We take a first step in this direction, focusing on models of guilt aversion in the Trust Game. We consider two alternative modeling assumptions: (i) guilt aversion depends on the role played in the game, because only the "trustee" can feel guilt for letting the co-player down, (ii) guilt aversion is independent of the role played in the game. We show how the set of Bayesian equilibria changes as the upper bound on guilt sensitivity varies, and we compare this with the complete-information case. Our analysis illustrates the incomplete-information approach to psychological games and can help organize experimental results in the Trust Game.

JEL classification: C72, C91, D03.

Keywords: Psychological games, Trust Game, guilt, incomplete information.

1 Introduction

The **Trust Game** is a stylized social dilemma whereby player A takes a costly action that generates a social return, and player B decides how to distribute the proceeds between himself and A. Experimental work on the Trust Game has shown systematic and significant departures from the standard equilibrium prediction implied by the assumption of common knowledge of selfish preferences (see Berg *et al.* 1995, Buskens & Raub 2008, Section III.A of the survey by Cooper & Kagel 2013, and the references therein). Given the simplicity of this game, such deviations are hard to explain as the result of bounded rationality. A recent paper by Charness & Dufwenberg (2006) provides support for the hypothesis that the behavior of most subjects in the second-mover role (B) is affected by aversion to letting down the first mover (A) relative to his expectations, as in Dufwenberg's (2002) model of marital investment. This is an instance of the "simple guilt" model of belief-dependent preferences of Battigalli & Dufwenberg (2007). Recent experimental work confirms this hypothesis (e.g., Reuben *et. al.* 2009, Chang *et al.* 2011, Bellemare *et al.* 2011, Attanasi *et al.* 2012).[1] Of course, when subjects' preferences

*Giuseppe Attanasi gratefully acknowledges financial support from ERC starting grant DU 283953. Pierpaolo Battigalli gratefully acknowledges financial support from ERC advanced grant 324219. Elena Manzoni gratefully acknowledges financial support from PRIN 2010-2011 "New Approaches to political economy: positive political theories, empirical evidence and experiments in laboratory". Moreover, authors thank Martin Dufwenberg, Nicodemo De Vito, Astrid Gamba, Marco Scarsini, Severine Toussaert and participants in the Sintelnet workshop in Toulouse and in the MMRG seminar series at the Catholic University of Milan for helpful comments.

[1] See also Dufwenberg & Gneezy (2000) and Guerra & Zizzo (2004). Vanberg (2008) and Ellingsen *et al.* (2010) question the guilt-aversion interpretation of pro-social behavior in the Trust Game.

differ from the simple benchmark of selfish expected payoff maximization, the assumption that such preferences are common knowledge is farfetched. Therefore, it should be assumed that the game played in the lab is one with *incomplete information*, even though the rules of the game (who plays when, information about previous moves and monetary payoffs at terminal nodes) are made common knowledge in experiments. This is consistent with the high heterogeneity of behavior and beliefs found in most experiments on other-regarding preferences (see Cooper & Kagel 2013). Our goal is to understand how such a game is played with incomplete information about guilt sensitivity.

We analyze two incomplete-information models of guilt aversion in the **Trust Minigame**, a binary-choice version of the Trust Game similar to the one analyzed by Charness & Dufwenberg (2006).[2] In the simpler model it is common knowledge that player A, the "truster", is selfish and only player B, the "trustee", can feel guilt. In the more complex model guilt sensitivity is not role-dependent. The first model is more tractable and it may be appropriate in situations where the players come from different populations. The second model may be more appropriate for situations where the subjects playing in roles A and B are drawn from the same population, as in most experiments. However, even when players are drawn from the same population, it is not implausible to assume that sensitivity to guilt is triggered only when playing in role B. This assumption is consistent with insights from the evolutionary psychology of emotions (Haselton & Ketelaar 2006).

In our analysis, as in the seminal paper by Harsanyi (1967-68), it is important to distinguish between two components of a player's type, the payoff type (here the sensitivity to guilt) and the epistemic type, which only determines beliefs. Hence, we allow for multiple types with the same guilt sensitivity and we obtain Bayesian equilibria with heterogeneous choices and heterogeneous (higher-order) beliefs about choices. This accords well with the experimental evidence cited above. Specifically, we adopt a simple and quite natural ordered parametrization of hierarchical beliefs with the following features. There are two possible guilt types of B, low and high. The epistemic type of A is parametrized by the subjective probability assigned by A to the high-guilt type of B; furthermore, A believes that guilt component and the epistemic component of B's type are independent and all types of A agree about the epistemic type of B. In the second model, where both A and B can feel guilt, similar assumptions hold for the beliefs of B about the type of A. In the first model, where it is common knowledge that A is selfish, the epistemic type of B parametrizes B's belief about the probability assigned by A to the high-guilt type of B, the higher B's epistemic type, the higher (in a stochastic order sense) such belief. Thus, in the parlance of incomplete information models, the epistemic type of A is parametrized by the exogenous first-order beliefs of A,[3] and the same holds for B in the model with role-independent guilt aversion. On the other hand, the epistemic type of B is parametrized by his exogenous second-order beliefs in the model where A is known to be selfish.

With this, in the model with role-dependent guilt, we show that in Bayesian equilibria where a positive fraction of A-types trust player B, higher types choose more pro-social actions and hold more optimistic hierarchical beliefs about actions. If the upper bound on guilt aversion is sufficiently high, there is only one equilibrium of this kind: Since the trusting action of (selfish) player A reveals A's hope that B will share, the high-guilt types of B choose this prosocial action independently of the epistemic component; thus the exogenous first-order belief of A about the guilt type of B coincides with the endogenous first-order belief about the prosocial choice of B. The maps from types to choices for both players are therefore determined by a kind a forward-induction argument. When instead the upper bound on guilt aversion is low, the epistemic component of B's type matters and the propensity to share is higher for higher epistemic types.

[2] We coined the name "Trust Minigame" after the "Ultimatum Minigame" of Binmore et al. (1995), a binary choice version of the Ultimatum Game.

[3] In Bayesian equlibrium analysis, beliefs about parameters are exogenous, beliefs about choice are endogenous. The same terminology applies to higher-order beliefs. See Section 3.

Furthermore, there may be multiple equilibria.

As explained above, the model with role-independent guilt is not a generalization of the previous one. In this case A's guilt sensitivity is not known and we assume that different types of B hold different beliefs about it, with higher types of B believing that A is more likely to be highly guilt averse. This yields a difference with respect to the previous model: the endogenous second-order belief that player B holds, if trusted, about A's belief that B would share is decreasing in B's epistemic type, hence B's choice is less pro-social for higher epistemic types. This is not surprising because here the meaning of the order on B's epistemic types is different from the previous model. Intuitively, the more B believes that A is guilt averse, the less he interprets A's trusting choice as pursuing a high material payoff, the less he is afraid that the selfish choice would disappoint A.

Related literature Our model finds its intellectual home in the theory of psychological games, that is, the analysis of games with belief-dependent preferences (Geanakoplos *et al.* 1989, Battigalli & Dufwenberg 2009, see also the introductory surveys by Dufwenberg 2006 and Attanasi & Nagel 2008). To our knowledge, this is one of the very few papers analyzing a psychological game with incomplete information, and the only one with a Bayesian equilibrium analysis of guilt aversion. Some papers analyze incomplete information models of games with belief-dependent preferences different from guilt aversion. Tadelis (2011) puts forward and validates experimentally a model of the Trust Minigame with incomplete information about player B's sensitivity to "shame". Caplin & Leahy (2004) analyze a model where a caring doctor has to decide whether to disclose health information to a patient with unknown propensity to anxiety. None of these models features heterogenous beliefs. Battigalli *et al.* (2012) analyze the cheap talk game of Gneezy's (2005) experiment under the assumption that the sender is affected by an unknown sensitivity to guilt. They show that, under mild and reasonable assumptions about the heterogeneous second-order beliefs of subjects playing in the role of the sender, guilt aversion explains the central tendencies of Gneezy's data on deception. Interestingly, they are able to derive such results without relying on Bayesian equilibrium analysis. Finally, our paper is related to Attanasi *et al.* (2012), who analyze experimentally the belief-dependent preferences, behavior and beliefs of subjects in the Trust Minigame. They show that making the elicited belief-dependent preferences common knowledge between the subjects of each matched pair significantly affects behavior and beliefs. This can be interpreted as comparison between a psychological game with incomplete information (control) and a psychological game with complete information (treatment). The theoretical comparison between treatment and control draws on the analysis of our paper, which therefore helps organizing the data of their experiment. More generally, we hope that our paper may have a pedagogical value for applied theorists and experimental economists who are interested in using psychological game theory to analyze social dilemmas.

The rest of the paper is structured as follows. Section 2 introduces the Trust Minigame with guilt aversion. Section 3 provides the methodology to analyze psychological Bayesian games, with a focus on the Trust Minigame with unknown guilt aversion. Section 4 puts forward and analyzes the model with role-dependent guilt, where A is known to be selfish. Section 5 puts forward and analyzes the model with role-independent guilt. Section 6 concludes. Formal proofs are collected in the Appendix.

2 Guilt aversion in the Trust Minigame

We analyze models of the Trust Minigame where players have different sensitivities to guilt feelings and incomplete information about the guilt sensitivity of the co-player. All the models we consider are based on the game form with monetary payoffs depicted in Figure 1.

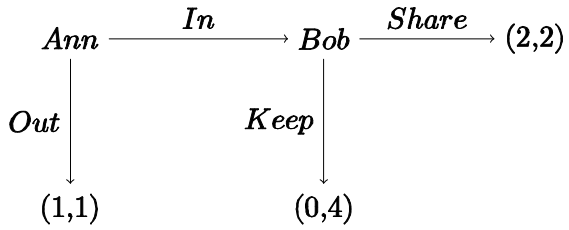

Figure 1. The Trust Minigame with material payoffs

In the analysis of this game form, we denote players' strategies as follows:

Strategy	*Notation*
In	I
Out	O
Share if *In*	S
Keep if *In*	K

In order to investigate the effects that the guilt feelings experienced by the players may have on their behavior, we need to consider their first and second-order beliefs about strategies. We denote with α_i player i's first-order beliefs, and with β_i the second-order beliefs.[4] Specifically, we use the notation described in the following table:[5]

Belief	*Notation*	*Definition*
Ann's initial **first-order belief**	α_A	$\mathbb{P}_A[S]$
Bob's initial **first-order belief**	α_B	$\mathbb{P}_B[I]$
A feature of Bob's initial **second-order belief**	$\beta_B^{\varnothing}$	$\mathbb{E}_B[\boldsymbol{\alpha}_A]$
A feature of Bob's **conditional second-order belief**	β_B^{I}	$\mathbb{E}_B[\boldsymbol{\alpha}_A \vert I]$

Note that we distinguish between the initial and conditional second-order beliefs of Bob, and we refer to the features of such beliefs that are relevant in our analysis. Indeed, we assume below that Bob's choice depends on his expectation of Ann's disappointment if he Keeps, which can be written as a function of the expected value of Ann's first-order belief. The second-order beliefs of Ann will be introduced later as needed.

According to the model of simple guilt (Battigalli & Dufwenberg, 2007), player i suffers from guilt to the extent that he believes that he is letting the co-player $-i$ down. In particular, player i has belief-dependent preferences over monetary payoff distributions represented by the following psychological utility function

$$u_i = m_i - \theta_i \max\{0, \mathbb{E}_{-i}[\mathbf{m}_{-i}] - m_{-i}\},$$

where $\theta_i \geq 0$ is the guilt sensitivity of i and $\max\{0, \mathbb{E}_{-i}[\mathbf{m}_{-i}] - m_{-i}\}$ measures the extent of the co-player's disappointment given his subjective beliefs.

We first assume that guilt sensitivity is **role-dependent**: only the second mover can be affected by guilt ($\theta_A = 0$, $\theta_B \geq 0$), and this is common knowledge. Ignoring players' beliefs about parameters, the strategic situation can be represented with the following parametrized psychological game:

[4] Because α and β are the first and second letter of the Greek alphabet.

[5] We use **bold** symbols to denote random variables. Since B does not know α_A, this number is a random variable from B's point of view, and its expectation is $\mathbb{E}_B[\boldsymbol{\alpha}_A]$. Similarly, we write $\mathbb{E}_A[\mathbf{m}_A]$ for the expected monetary payoff of A.

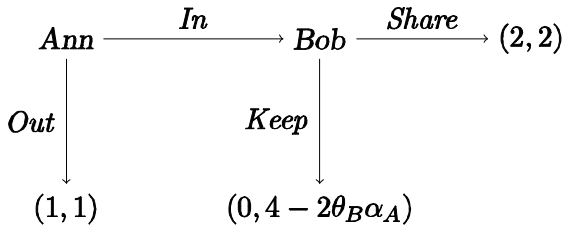

Figure 2. The Trust Minigame with psychological utilities

Indeed, Ann can only be disappointed after terminal history (I, K), in which case the extent of her disappointment is

$$\max\{0, \mathbb{E}_A[\mathbf{m}_A] - m_A(I, K)\} = 2 \cdot \alpha_A + 0 \cdot (1 - \alpha_A) = 2\alpha_A,$$

where $m_i(z)$ denotes the material payoff of i at terminal history z. Thus, the psychological utility of $z = (I, K)$ for Bob (expressed as a function of Ann's first-order belief α_A) is

$$u_B(I, K, \alpha_A) = m_B(I, K) - \theta_B \max\{0, \mathbb{E}_A[\mathbf{m}_A] - m_A(I, K)\} = 4 - 2\theta_B\alpha_A.$$

Of course, when Bob evaluates his alternatives and chooses his optimal strategy, he compares the utility from choosing S with the expected psychological utility from choosing K, which depends on his second-order beliefs. As long as Bob initially assigns a strictly positive probability to I ($\alpha_B = \mathbb{P}_B[I] > 0$) the comparison between strategy S and K can equivalently be made either *ex ante*, or conditional on I, because the difference between the ex ante expected utilities of S and K is proportional to the difference between the conditional expected utilities of S and K:

$$\mathbb{E}_B^S[\mathbf{u}_B] - \mathbb{E}_B^K[\mathbf{u}_B] = \mathbb{E}_B[u_B(I, S, \boldsymbol{\alpha}_A) - u_B(I, K, \boldsymbol{\alpha}_A)|I] \cdot \mathbb{P}_B[I] = [2 - (4 - 2\theta_B\beta_B^I)] \cdot \alpha_B.$$

By definition, $0 \leq \beta_B^I \leq 1$, thus Bob Shares if $\theta_B > 1$ and $\beta_B^I > \frac{1}{\theta_B}$.

The assumption that guilt sensitivity depends on one's role in a game is consistent with insights from the evolutionary psychology of emotions, which suggests that when a single emotion operates in a variety of different domains its effects are moderated by contextual cues (Haselton & Ketelaar 2006). Since this assumption simplifies the analysis, we maintain it in the first part of the paper. In Section 5 we analyze a model where guilt sensitivity is role-independent.

In all the models considered below, for all parameter values, there is an equilibrium where Ann goes Out with probability one: If Ann is certain that Bob would Keep ($\alpha_A = 0$), she stays Out; if Bob beliefs are correct, $\alpha_B = 0$ and $\beta_B^{\varnothing} = 0$; then β_B^I is not pinned down by Bayes rule, but as long as $\beta_B^I < \frac{1}{\theta_B}$ Bob's optimal strategy is indeed to Keep, exactly what Ann expects.[6]

Yet, casual evidence and the experimental evidence cited in the Introduction show that positive fractions of agents systematically trust co-players and share with co-players**.** Therefore, when O is not the unique equilibrium outcome of the model, we focus on the more interesting equilibria where trust and sharing occur with positive probability. We call such equilibria **"non-trivial"** because they are the equilibria where guilt aversion plays a role.

3 Methodology: Bayesian psychological games

We are going to model incomplete information about θ using the methodology first proposed by Harsanyi (1967-68), suitably extended to psychological games (see also Battigalli & Dufwenberg

[6] This is an instance of the following observation (cf. Battigalli & Dufwenberg 2007, Observation 2). Fix a game form with material payoffs and no chance moves. Let G denote the corresponding complete information game obtained when the game form and the fact that players are selfish are common knowledge. Let $\Gamma(G)$ denote any psychological game obtained from G by adding to each player's material payoff a guilt-aversion term and possibly allowing incomplete information about guilt parameters. Then every pure strategy (sequential) equilibrium of the material-payoffs game G is also a (Bayesian perfect) equilibrium of the psychological game $\Gamma(G)$.

2009, Section 6.2). We define type structures that implicitly determine the possible hierarchies of subjective beliefs of the players.

Although our methodology is fully standard from the abstract theory perspective, it is not widely used in applied theory. Therefore, it is useful to describe carefully the building blocks of our approach.

A note on terminology We call **"exogenous"** a belief about an exogenous variable or a parameter: a belief about θ is an exogenous first-order belief, a joint belief about θ and exogenous first-order beliefs of the co-player is an exogenous second-order belief, and so on. We call **"endogenous"** a belief about a variable that we try to explain, or predict, with the strategic analysis of the game. In particular, a belief about strategies is an endogenous first-order belief, a joint belief about strategies and endogenous first-order beliefs is an endogenous second-order belief, and so on. We also call "endogenous" a joint belief about exogenous and endogenous variables.[7] Also, we distinguish between "player", which corresponds to the role (A or B) in the game, and the individual playing in role A or B, whom we call "agent". An agent is equivalently called "subject" when we refer to implementations of the game in laboratory experiments.

3.1 Type structures and hierarchies of beliefs

We consider situations where the psychological utility functions of players A and B are determined by parameters $\theta_A \in \Theta_A$, $\theta_B \in \Theta_B$ known to A and B respectively, called the utility types of A and B. Formally, psychological utility is a parametrized function[8]

$$u_i : \Theta_i \times Z \times \mathcal{H}_i \times \mathcal{H}_{-i} \to \mathbb{R}$$

where Z is the set of terminal histories of the game and $\mathcal{H}_i$ ($\mathcal{H}_{-i}$) is a space of endogenous hierarchical beliefs of player i ($-i$).[9] Since in our applications θ_i is the guilt sensitivity parameter of player i, we call θ_i a **guilt type**. When the parameter set of player i, Θ_i, is a singleton, the guilt type of i is common knowledge. In models with role-dependent guilt sensitivity we have $\Theta_A \neq \Theta_B$; in particular, we assume that Θ_A is a singleton, because player A is commonly known to be a selfish expected material payoff maximizer. In models with role-independent guilt sensitivity $\Theta_A = \Theta_B = \Theta$. In all our models we assume that $\Theta_B = \{\theta^L, \theta^H\}$ with $0 = \theta^L < \theta^H$. This simplifies the parametrization of beliefs.[10] Our analysis can be extended to the case where Θ_i is an interval.

The subjective exogenous beliefs of A and B about each other private information and beliefs are implicitly represented by a **type structure**, that is, a tuple

$$\mathcal{T} = \langle I = \{A, B\}, (\Theta_i, T_i, \boldsymbol{\vartheta}_i : T_i \to \Theta_i, \boldsymbol{\tau}_i : T_i \to \Delta(T_{-i}))_{i \in I} \rangle .$$

Elements of T_i are called Harsanyi types, or simply **types**. An Harsanyi type specifies both the guilt type (more generally the utility type) and the exogenous beliefs of player i. The following are technical assumptions: for each player i, T_i is a *compact metric* space, the set of Borel probability measures $\Delta(T_{-i})$ is endowed with the topology of weak convergence (hence it is compact and metrizable), and the functions $\boldsymbol{\vartheta}_i(\cdot)$, $\boldsymbol{\tau}_i(\cdot)$ are *continuous*. Also note that we use bold letters to

[7] This terminology is appropriate because we are *not* trying to analyze stationary states of learning dynamics. If this instead were the case, we would use an appropriate version of the self-confirming (or conjectural) equilibrium concept, and beliefs about θ would be part of what is to be explained,. i.e. they would be "endogenous" as well (see Esponda, 2012).

[8] Since θ_i is just a preference parameter, it is appropriate to assume *private values*: the utility of i does not depend on θ_{-i}.

[9] See Geanakoplos *et al.* (1989) and Battigalli & Dufwenberg (2009). In the latter, $\mathcal{H}_i$ is a space of hierarchical *conditional* beliefs. It can be shown that utility function u_i can be replaced by a utility function $\bar{u}_i : \Theta_i \times Z \times \mathcal{H}_i \to \mathbb{R}$ inducing the same best reply correspondence, that depends only on the endogenous beliefs of i. This observation is used in the representation of the Trust mini-Game with guilt aversion in Section 5, Figure 3.

[10] The analysis of the model with role-dependent guilt gives the same results whenever $0 \leq \theta^L < 1$.

denote functions interpreted as **random variables**, that is, functions that depend on the **state of the world** (t_A, t_B). Function $\boldsymbol{\vartheta}_i(\cdot)$ specifies the psychological utility (guilt sensitivity) of type t_i, and function $\boldsymbol{\tau}_i(\cdot)$ determines the beliefs of t_i about the utility and beliefs of the co-player $-i$. In particular, the type structure yields, for each type of each player, an implicit description of hierarchical exogenous beliefs, as explained below. Given a random variable $\mathbf{x}_i : T_i \to X_i$, we denote events about $\mathbf{x}_i$ either directly as subsets of T_i, or according to the convention which is common in statistics. For example, both $\boldsymbol{\vartheta}_i^{-1}(\theta^H)$ and $\boldsymbol{\vartheta}_i = \theta^H$ denote the event that the guilt type of i is θ^H. We use whatever notation is more convenient and transparent in the given context.

Once we append a type structure to the profile of parametrized utility functions, we obtain a **Bayesian psychological game**:

$$\Gamma = \langle I = \{A, B\}, (\Theta_i, u_i : \Theta_i \times Z \times \mathcal{H}_i \times \mathcal{H}_{-i} \to \mathbb{R}, T_i, \boldsymbol{\vartheta}_i : T_i \to \Theta_i, \boldsymbol{\tau}_i : T_i \to \Delta(T_{-i}))_{i \in I} \rangle$$

3.1.1 Epistemic types

We will consider type structures where the set of types T_i can be factorized as $T_i = \Theta_i \times T_i^e$ where T_i^e is a set T_i^e of **epistemic types**. The latter parametrize exogenous beliefs and we assume that the parameter space is $T_i^e = [0, 1]$. Therefore Harsanyi types are pairs given by a guilt type and an epistemic type: $t_i = (\theta_i, e_i) \in \Theta_i \times [0, 1] = T_i$. Then function $\boldsymbol{\vartheta}_i : \Theta_i \times T_i^e \to \Theta_i$ is just the projection of $T_i = \Theta_i \times [0, 1]$ onto Θ_i (that is, $\boldsymbol{\vartheta}_i(\theta_i, e_i) = \theta_i$ for each (θ_i, e_i)). Furthermore, we assume that $\boldsymbol{\tau}_i(\theta_i, e_i)$ *depends only on* e_i and is *monotone:* roughly, higher epistemic types of player i assign higher probability to high guilt and/or epistemic types of the co-player $-i$.[11]

3.1.2 Exogenous n-th order beliefs

The **exogenous first-order belief** of a type t_i is determined by the equation

$$\mathbf{p}_i^1(t_i)[E_{-i}^0] = \boldsymbol{\tau}_i(t_i)[(\boldsymbol{\vartheta}_{-i})^{-1}(E_{-i}^0)] \ (E_{-i}^0 \subseteq \Theta_{-i} \text{ Borel measurable}).$$

This way we obtain a map $(\boldsymbol{\vartheta}_i, \mathbf{p}_i^1) : T_i \to \Theta_i \times \Delta(\Theta_{-i})$ for each $i \in \{A, B\}$. Then the **exogenous second-order belief** of a type t_i is determined by the equation

$$\mathbf{p}_i^2(t_i)[E_{-i}^1] = \boldsymbol{\tau}_i(t_i)[(\boldsymbol{\vartheta}_{-i}, \mathbf{p}_{-i}^1)^{-1}(E_{-i}^1)] \ (E_{-i}^1 \subseteq \Theta_{-i} \times \Delta(\Theta_i) \text{ Borel measurable}).$$

Proceeding this way, we can associate a **hierarchy of exogenous beliefs** with each type. However, beliefs beyond the second-order will not be used in the analysis below.

3.2 Equilibrium

A **Bayesian equilibrium** of the psychological Trust Minigame with incomplete information is given by a pair of measurable decision functions $(\boldsymbol{\sigma}_A : T_A \to \{I, O\}, \boldsymbol{\sigma}_B : T_B \to \{S, K\})$ such that for each player $i \in \{A, B\}$ and type $t_i \in T_i$, choice $\boldsymbol{\sigma}_i(t_i)$ maximizes i's expected psychological utility, given the endogenous beliefs of type t_i about the co-player's choice and beliefs.[12] In a **perfect** Bayesian equilibrium, player B maximizes his *conditional* expected utility upon observing I, with conditional beliefs computed by Bayes rule, if possible. However, in the non-trivial equilibria we are going to focus on, action I is chosen by a positive fraction of types,

[11] If instead τ_i also depends monotonically on θ_i, we have a form of perception of false consensus, that is, player $-i$ believes that higher guilt types of i are associated with higher beliefs about the guilt type of $-i$. See Section 6 for a discussion of false consensus.

[12] Such decision functions are often called "strategies". We avoid this terminology for two reasons. First, we are not studying a situation where player i decides how to play the game before being informed about his type; rather we study decisions of different agents playing in role i, where each agent is characterized by some type t_i. Second, we want to avoid confusion with the strategies of the Trust Minigame, such as "Share if In".

hence it has positive probability. As we noticed in Section 2, in this case *ex ante* maximization of psychological utility is equivalent to conditional maximization; therefore *non-trivial Bayesian equilibria are also perfect.*

3.2.1 Endogenous n-th order beliefs and other random variables

It is important to understand how the type structure and decision functions $\boldsymbol{\sigma}_i$ generate the endogenous beliefs of the players. We analyze psychological games where the utility of i (determined by his guilt type θ_i) depends on the strategy of $-i$ (identified with $-i$'s plan) and on the first-order endogenous beliefs of $-i$. For example, the utility of guilt type θ_B depends on B's material payoff determined by the sequence of actions and on the disappointment of A; the latter is positive if A plans to choose I, carries out such plan and then B replies with K, in this case A's disappointment is determined by the first-order belief of A about the choice of B, that is, the probability α_A assigned by A to strategy S.

The latter probability is an endogenous first-order belief determined by the type of A and the equilibrium decision function of B:

$$\boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B^{-1}(S)] = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B = S]. \tag{1}$$

For player B (and the analyst), $\boldsymbol{\alpha}_A : T_i \to [0,1]$ is a random variable. Player B can compute his initial expectation of $\boldsymbol{\alpha}_A$ as follows:[13]

$$\boldsymbol{\beta}_B^{\varnothing}(t_B) = \mathbb{E}_{t_B}[\boldsymbol{\alpha}_A] = \int \boldsymbol{\alpha}_A(t_A)\boldsymbol{\tau}_B(t_B)[\mathrm{d}t_A]. \tag{2}$$

Since B takes an action only if he observes I, his choice depends on his second-order belief conditional on I:[14]

$$\boldsymbol{\beta}_B^I(t_B) = \mathbb{E}_{t_B}[\boldsymbol{\alpha}_A|\boldsymbol{\sigma}_A = I] = \frac{\int_{(\boldsymbol{\sigma}_A)^{-1}(I)} \boldsymbol{\alpha}_A(t_A)\boldsymbol{\tau}_B(t_B)[\mathrm{d}t_A]}{\boldsymbol{\tau}_B(t_B)[\boldsymbol{\sigma}_A = I]}, \text{ if } \boldsymbol{\tau}_B(t_B)[\boldsymbol{\sigma}_A = I] > 0. \tag{3}$$

As the above equations illustrate, all the endogenous beliefs are implicitly determined by the equilibrium decision functions $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ (given the type structure). However, for the sake of clarity, in our analysis we will make the key endogenous beliefs explicit.

Beside first and second-order endogenous beliefs, the type structure and decision functions determine other random variables that will be used in our analysis (all written in bold). For example, the random variable "monetary payoff of player i" is[15]

$$\mathbf{m}_i(t_A, t_B) = \begin{cases} m_i(O), & \text{if } \boldsymbol{\sigma}_A(t_A) = O, \\ m_i(I,K), & \text{if } \boldsymbol{\sigma}_A(t_A) = I,\ \boldsymbol{\sigma}_B(t_B) = K, \\ m_i(I,S), & \text{if } \boldsymbol{\sigma}_A(t_A) = I,\ \boldsymbol{\sigma}_B(t_B) = S, \end{cases}$$

and the random variable "psychological utility of player i" is

$$\mathbf{u}_i(t_A, t_B) = \mathbf{m}_i(t_A, t_B) - \boldsymbol{\vartheta}_i(t_i)\max\{0, \mathbb{E}_{t_{-i}}[\mathbf{m}_{-i}] - \mathbf{m}_{-i}(t_A, t_B)\},$$

where, of course, in the computation of $\mathbb{E}_{t_{-i}}[\mathbf{m}_{-i}]$ type t_{-i} assigns probability one to the choice $\boldsymbol{\sigma}_{-i}(t_{-i})$.

Furthermore, the epistemic type of player i is a random variable from the point of view of the co-player $-i$. Formally, this random variable is just the projection from $T_A \times T_B$ onto T_i^e: $\mathbf{e}_i(t_A, t_B) = e_i$ if and only if $t_i = (\theta_i, e_i)$ for some $\theta_i \in \Theta_i$. Thus, for example, $[\mathbf{e}_i > x]$ denotes the event that the epistemic type of i is higher than x.

[13] Given a real-valued random variable $\mathbf{x}_{-i} : T_{-i} \to \mathbb{R}$ and a measure $\mu \in \Delta(T_{-i})$, $\mathbb{E}_\mu[\mathbf{x}_{-i}]$ denotes the expectation of $\mathbf{x}_{-i}$ according to μ. To ease notation for the expectation of $\mathbf{x}_{-i}$ according to the belief of type t_i, we write $\mathbb{E}_{t_i}[\mathbf{x}_{-i}]$ instead of $\mathbb{E}_{\boldsymbol{\tau}_i(t_i)}[\mathbf{x}_{-i}]$.

[14] Recall that $\boldsymbol{\tau}_B(t_B)[\boldsymbol{\sigma}_A = I] > 0$ for every non-trivial equilibrium.

[15] Recall that $m_i(z)$ is the monetary payoff of player i at terminal history z.

3.2.2 Actual distributions

In this paper we focus on the equilibrium derivation of the decision functions and of other endogenous random variables specified above. Therefore, it is not necessary in our analysis to postulate an objective statistical distribution on the type space.[16] Of course, such a distribution is necessary to obtain statistical predictions from equilibrium analysis. We extensively comment on this in Section 6.

4 Role-dependent guilt

We start our analysis with a simple model where player A is commonly known to be selfish, i.e. a monetary payoff maximizer. For example, there may be an heterogeneous population of individuals from which players are drawn at random and assigned to roles A and B. The "potential" guilt sensitivity θ in this population can be either high, $\theta^H > 0$, or low, $\theta^L = 0$. But *the actual guilt sensitivity is triggered by the role in the game: only player B can feel guilt.*

Alternatively, we can imagine that it is commonly known that the individual playing in role A is drawn from a population of selfish agents, whereas the individual playing in role B is drawn from an heterogeneous population where some agents are guilt averse. Of course, this second interpretation does not fit well with standard experimental implementations of the Trust Minigame.

4.1 Complete Information model

Before we move on to the details of the incomplete information model, it is useful to report the results about the complete information benchmark (cf. Dufwenberg 2002). The psychological game is described in Figure 2 of Section 2, assuming that $\theta_B > 0$ is commonly known.[17]

As explained in Section 2, player B Shares (respectively Keeps) if his conditional second-order belief satisfies $\beta_B^I > \frac{1}{\theta_B}$ (respectively $\beta_B^I < \frac{1}{\theta_B}$). In a complete information equilibrium first and second-order beliefs are correct, and players maximize given their belief. Therefore, for each $\theta_B > 0$, the profile $(O, K, \alpha_A, \beta_B^{\varnothing}, \beta_B^I)$ is an equilibrium if $\alpha_A = \beta_B^{\varnothing} = 0$ and $\beta_B^I < \frac{1}{\theta_B}$. Intuitively, when player A thinks that B will Keep, A's optimal choice is to go Out. If B initially expects this ($\alpha_B = 0$), the conditional second-order belief β_B^I is not pinned down by Bayes rule. For any out of equilibrium belief $\beta_B^I < \frac{1}{\theta_B}$, B's optimal choice is to Keep. This is an instance of the general observation made at the end of Section 2: the equilibrium (O, K) of the material payoff game with selfish preferences is also an equilibrium of the psychological games with guilt aversion analyzed in this paper. Recall that $\beta_B^I = \mathbb{E}_B[\boldsymbol{\alpha}_A | I] \in [0, 1]$. Therefore, if $\theta_B \in (0, 1)$, then $\beta_B^I < \frac{1}{\theta_B}$ and this is the unique equilibrium of the complete information game with guilt aversion.

For higher values of B's type, $\theta_B \in [1, +\infty)$, there is another pure strategy equilibrium, $(I, S, \alpha_A = \beta_B^{\varnothing} = \beta_B^I = 1)$. To see this, notice that if A expects B to Share with probability $\alpha_A \geq \frac{1}{2}$ (and in particular with probability $\alpha_A = 1$) it is optimal for A to go In. Moreover, if $\beta_B^I \geq \frac{1}{\theta_B}$ (in particular if $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = 1$) it is optimal for B to Share when A goes In. This is also the unique *forward induction* equilibrium for $\theta_B \in (2, +\infty)$. Indeed, A finds it optimal to go In if and only if $\alpha_A \geq \frac{1}{2}$. Therefore, if B rationalizes A's observed choice of going In, his conditional second-order belief satisfies $\beta_B^I \geq \frac{1}{2}$, which implies that it is (uniquely) optimal to Share, as $\beta_B^I \geq \frac{1}{2} > \frac{1}{\theta_B}$. In Appendix A.1 we provide a complete description of the equilibrium correspondence, including mixed equilibria.

[16] As the reader may have noticed, we have *not* assumed that the functions associating each type with an exogenous belief are derived from an "objective" common prior on the state space.

[17] When it is common knowledge that $\theta_B = 0$ the analysis is trivial, as we obtain the usual backward induction equilbrium with selfish preferences.

4.2 A model with incomplete information and heterogeneous beliefs

4.2.1 Type structure

We have a continuum of types on both sides. The beliefs of player i about the type of the co-player $-i$ are determined by an epistemic parameter $e_i \in [0,1]$. We assume for simplicity that each type of player A believes that the guilt and epistemic type of B are statistically independent. Since the guilt type of A is commonly known to be zero, for player A Harsanyi types and epistemic types coincide: $T_A = T_A^e = [0,1]$ (in our more general formalism, $T_A = \{0\} \times T_A^e$, which is isomorphic to T_A^e). We let $e_A = t_A \in T_A$ parametrize the subjective probability of the high-guilt type of B: $t_A = \mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B = \theta^H]$. All types of A have common marginal beliefs about the epistemic type of B given by a *continuous* cdf $F : \mathbb{R} \to [0,1]$ with *support* $[0,1]$ (that is, strictly increasing on $[0,1]$ with $F(0) = 1 - F(1) = 0$). Each epistemic type e_B of B has a belief about A's type given by a *continuous* cdf $F_{e_B} : \mathbb{R} \to [0,1]$ with *support* $[0,1]$.

Specifically, we let $T_A = [0,1]$, $T_B = \{\theta^L, \theta^H\} \times [0,1]$, $\boldsymbol{\tau}_i : T_i \to \Delta(T_{-i})$ $(i \in \{A,B\})$ with

$$\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H \cap \mathbf{e}_B \leq y] = t_A F(y), \tag{4}$$

and

$$\boldsymbol{\tau}_B(\theta_B, e_B)[\mathbf{t}_A \leq x] = F_{e_B}(x), \tag{5}$$

for all $t_A, e_B, x, y \in [0,1]$, $\theta_B \in \{\theta^L, \theta^H\}$.

According to our general assumptions about type structures, the map $e_B \longmapsto F_{e_B}(\cdot)$ is continuous.[18] Furthermore, we assume that the following stochastic order property holds: the conditional expectations $\mathbb{E}_{e_B}[\mathbf{t}_A | \mathbf{t}_A > x]$ are strictly increasing in e_B, that is,

$$e_B < \bar{e}_B \Rightarrow \frac{1}{1 - F_{e_B}(x)} \int_x^1 t_A \mathrm{d}F_{e_B}(t_A) < \frac{1}{1 - F_{\bar{e}_B}(x)} \int_x^1 t_A \mathrm{d}F_{\bar{e}_B}(t_A) \tag{6}$$

for all $x \in [0,1)$ and $e_B, \bar{e}_B \in [0,1]$. Intuitively, this means that higher epistemic types of B have higher beliefs about the (epistemic) type of A,[19] as anticipated in Section 3.1. All of the above is common knowledge. Since the beliefs of type (θ_B, e_B) depend only on the epistemic component e_B, to ease notation we write $\boldsymbol{\alpha}_B(e_B)$, $\boldsymbol{\beta}_B^{\varnothing}(e_B)$, $\boldsymbol{\beta}_B^I(e_B)$ instead of, respectively, $\boldsymbol{\alpha}_B(\theta_B, e_B)$, $\boldsymbol{\beta}_B^{\varnothing}(\theta_B, e_B)$, $\boldsymbol{\beta}_B^I(\theta_B, e_B)$.

4.2.2 Equilibrium analysis

As explained in Section 2, there is always an equilibrium outcome where all A-types stay Out. In this paper, we focus on the more interesting *non-trivial* equilibria where a positive fraction of A-types choose In. It turns out that all non-trivial equilibria exhibit threshold decision functions. Since beliefs are described by atomless distributions, the choice of the threshold type (who is indifferent) is immaterial for equilibrium analysis. We assume wlog that such type chooses the "low" action, that is, O for player A and K for player B. In what follows, we say that a decision function $\boldsymbol{\sigma}_A$ is **(weakly) increasing** if there is a threshold $\hat{e}_A \in (0,1)$ $(\hat{e}_A \in [0,1])$ such that

[18] We assumed that $\boldsymbol{\tau}_B : T_B \to \Delta(T_A)$ is continuous. In the present model this means that $e_B \to \bar{e}_B$ implies that $F_{e_B}(t_A) \to F_{\bar{e}_B}(t_A)$ for every continuity point of $F_{\bar{e}_B}$. Since $F_{\bar{e}_B}$ is assumed to be continuous, $F_{e_B}(\cdot)$ must converge to $F_{\bar{e}_B}(\cdot)$ pointwise.

[19] This assumption holds if the epistemic types of B are *ordered by hazard rate*. When every cdf F_{e_B} is differentiable, with $f_{e_B} = F'_{e_B}$, this can be expressed as follows:

$$e_B < \bar{e}_B \Rightarrow \frac{f_{\bar{e}_B}(t_A)}{1 - F_{\bar{e}_B}(t_A)} < \frac{f_{e_B}(t_A)}{1 - F_{e_B}(t_A)}$$

for all $e_B, \bar{e}_B \in [0,1]$ and $t_A \in [0,1)$. See Shaked and Shantikumar (2007, pp 16-17).

$\boldsymbol{\sigma}_A(t_A) = O$ iff (if and only if) $t_A \leq \hat{e}_A$. A similar terminology is used for the high-guilt type of B: $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ is **(weakly) increasing** if there is $\hat{e}_B^H \in (0,1)$ ($\hat{e}_B^H \in [0,1]$) such that $\boldsymbol{\sigma}_B(\theta^H, e_B) = K$ iff $e_B \leq \hat{e}_B^H$. In particular, a weakly increasing decision function is either increasing or essentially constant.[20]

A non-trivial equilibrium is given by decision functions $\boldsymbol{\sigma}_A$ and $\boldsymbol{\sigma}_B$, which in turn determine the endogenous belief functions $\boldsymbol{\alpha}_A$, $\boldsymbol{\beta}_B^{\varnothing}$ and $\boldsymbol{\beta}_B^I$ as in eq. (1)-(2)-(3), so that $\boldsymbol{\sigma}_A^{-1}(I)$ (a measurable subset of $[0,1]$) has positive measure, $\boldsymbol{\sigma}_A(t_A)$ is a best reply to $\boldsymbol{\alpha}_A(t_A)$ for all t_A, and $\boldsymbol{\sigma}_B(\theta_B, e_B)$ is a best reply to $\boldsymbol{\beta}_B^I(e_B)$ for all t_A, θ_B and e_B.

Proposition 1 *In the model given by (4)-(5)-(6) there is a non-trivial equilibrium only if $\theta^H > 1$. Every non-trivial equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ has the following structure: $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$ for every e_B, $\boldsymbol{\sigma}_A$ is increasing, $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ is weakly increasing and they are respectively characterized by thresholds $\hat{e}_A \in (0,1)$ and $\hat{e}_B^H \in [0,1)$ such that*

(a) *for every type t_A,*

$$\boldsymbol{\alpha}_A(t_A) = t_A\left(1 - F\left(\hat{e}_B^H\right)\right),$$

hence $\boldsymbol{\alpha}_A(\cdot)$ is strictly increasing and the incentive condition for A yields

$$\hat{e}_A = \frac{1}{2\left(1 - F\left(\hat{e}_B^H\right)\right)};$$

(b) *for every epistemic type e_B,*

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = \left(1 - F\left(\hat{e}_B^H\right)\right)\int_0^1 t_A \mathrm{d}F_{e_B}(t_A),$$

$$\frac{1}{2} < \boldsymbol{\beta}_B^I(e_B) = \frac{1 - F(\hat{e}_B^H)}{1 - F_{e_B}(\hat{e}_A)}\int_{\hat{e}_A}^1 t_A \mathrm{d}F_{e_B}(t_A) < 1,$$

hence, by assumption (6), $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ and $\boldsymbol{\beta}_B^I(\cdot)$ are strictly increasing and the incentive condition for B yields

$$\begin{aligned} \hat{e}_B^H &= 0 \Rightarrow \boldsymbol{\beta}_B^I(e_B) \geq \frac{1}{\theta^H}, \\ \hat{e}_B^H &> 0 \Rightarrow \boldsymbol{\beta}_B^I(e_B) = \frac{1}{\theta^H}. \end{aligned}$$

Sketch of proof First note that a low-guilt type of B always Keeps: $\boldsymbol{\sigma}_B(\theta_B, e_B) = K$ if $\theta_B < 1$. This gives the necessary condition for existence of non-trivial equilibria. Furthermore, in a non-trivial equilibrium A's first-order endogenous belief $\boldsymbol{\alpha}_A(t_A) = \mathbb{P}_{t_A}[\boldsymbol{\sigma}_B = S]$ is increasing in t_A: B plays S only if his guilt type is high, therefore $\boldsymbol{\alpha}_A(t_A)$ is the product of two probabilities, the probability that B's guilt type is high, $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B = \theta^H] = t_A$, and the probability $\mathbb{P}_F[\boldsymbol{\sigma}_B = S | \boldsymbol{\vartheta}_B = \theta^H]$ that B chooses S given that his guilt type is high. By eq. (4), the latter is determined by cdf F independently of t_A:

$$\boldsymbol{\alpha}_A(t_A) = t_A \mathbb{P}_F[\boldsymbol{\sigma}_B = S | \boldsymbol{\vartheta}_B = \theta^H].$$

[20] On the other hand, when we speak of a real valued function $\varphi : X \to \mathbb{R}$ with $X \subseteq \mathbb{R}$ (such as a cdf or a belief function), we say that φ is **strictly increasing** if $x' > x'' \Rightarrow \varphi(x') > \varphi(x'')$, and we just say that φ is **increasing** if $x' \geq x'' \Rightarrow \varphi(x') \geq \varphi(x'')$.

Since $\boldsymbol{\sigma}_A(t_A) = I$ iff $\boldsymbol{\alpha}_A(t_A) > \frac{1}{2}$, it follows that $\boldsymbol{\sigma}_A(t_A) = I$ iff $t_A > \hat{e}_A$, where $\boldsymbol{\alpha}_A(\hat{e}_A) = \frac{1}{2}$, that is

$$\hat{e}_A = \frac{1}{2\mathbb{P}_F[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H]}.$$

We now need to analyze B's equilibrium decision function when the guilt type of B is high, that is, $\boldsymbol{\sigma}_B(\theta^H, e_B)$. Since F_{e_B} has full support and a positive fraction of A-types chooses I, every epistemic type e_B assigns a strictly positive probability to the event "A chooses I": $\mathbb{P}_{e_B}[\mathbf{t}_A > \hat{e}_A] = 1 - F_{e_B}(\hat{e}_A) > 0$. Hence, B's second-order endogenous belief is determined by the equation $\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A|\mathbf{t}_A > \hat{e}_A]$. By assumption (6), $\mathbb{E}_{e_B}[\mathbf{t}_A|\mathbf{t}_A > \hat{e}_A]$ is increasing in e_B, higher epistemic types of B hold higher conditional beliefs about the (epistemic) type of A; this also implies that they hold higher second-order endogenous beliefs, because $\boldsymbol{\alpha}_A$ is an increasing linear function of t_A. Therefore B's second-order endogenous belief, $\boldsymbol{\beta}_B^I(e_B)$, is increasing in e_B and the decision function satisfies $\boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = S$ iff $e_B > \hat{e}_B^H$, for some $\hat{e}_B^H$. Thus $\mathbb{P}_F[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H] = 1 - F(\hat{e}_B^H)$ and

$$\boldsymbol{\alpha}_A(t_A) = t_A(1 - F(\hat{e}_B^H)),$$

which gives the formulas for $\boldsymbol{\beta}_B^{\varnothing}(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A]$ and $\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A|\mathbf{t}_A > \hat{e}_A]$. The indifference condition for a positive threshold $\hat{e}_B^H$ is $\boldsymbol{\beta}_B^I(\hat{e}_B^H) = \frac{1}{\theta^H}$. A more formal proof is contained in Appendix A.2.

4.2.3 High upper bound on guilt $\left(\theta^H > 2\right)$.

In the general case analyzed above, it is possible to have multiple non-trivial equilibria; it is also possible that non-trivial equilibria do not exist even if $\theta^H > 1$. We will show this in a parametric example (see Section 4.2.4). When instead B's upper bound on guilt sensitivity is sufficiently high ($\theta^H > 2$), we can show that a non-trivial equilibrium exists, it is unique and has a simple form. Intuitively, this is due to the fact that with a high θ^H we can apply a forward-induction argument: A chooses I iff $\boldsymbol{\alpha}_A > \frac{1}{2}$, thus I reveals $\boldsymbol{\alpha}_A > \frac{1}{2}$ to player B, which implies that his conditional second-order belief is also higher than $\frac{1}{2}$: $\boldsymbol{\beta}_B^I = \mathbb{E}[\boldsymbol{\alpha}_A|\boldsymbol{\alpha}_A > \frac{1}{2}] > \frac{1}{2}$. When $\theta^H > 2$, this implies that B chooses S *if* (and only if) his guilt type is high ($\boldsymbol{\vartheta}_B = \theta^H$). It follows that, for player A, the probability of S coincides with the probability of the high-guilt type, and A chooses I iff this exogenous probability is higher than $\frac{1}{2}$. This completely determines the non-trivial equilibrium. This argument holds without assuming the stochastic order condition (6). If this condition holds, then second-order endogenous beliefs are strictly increasing. With this, the formal proof of the following proposition is straightforward and hence we leave it to the reader.

Proposition 2 *In the model given by (4)-(5), if $\theta^H > 2$ there is a unique non-trivial Bayesian equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$, and it has the following properties:*

(a) *for every type t_A, $\boldsymbol{\sigma}_A(t_A) = O$ iff $t_A \leq \frac{1}{2}$ and $\boldsymbol{\alpha}_A(t_A) = t_A$;*

(b) *for every epistemic type e_B, $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$, $\boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = S$ and*

$$\begin{aligned}
\boldsymbol{\beta}_B^{\varnothing}(e_B) &= \int_0^1 t_A \mathrm{d}F_{e_B}(t_A), \\
\boldsymbol{\beta}_B^I(e_B) &= \left(1 - F_{e_B}\left(\frac{1}{2}\right)\right)^{-1} \int_{\frac{1}{2}}^1 t_A \mathrm{d}F_{e_B}(t_A) > \frac{1}{2};
\end{aligned}$$

(c) *if (6) holds $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ and $\boldsymbol{\beta}_B^I(\cdot)$ are strictly increasing.*

4.2.4 A parametric example

We analyze a parametric representation of the exogenous beliefs of the players to illustrate our modeling approach with a specific example. The type structure is specified as follows: all types of A have a uniform distribution on the epistemic types of B, and each epistemic type $e_B \in (0,1)$ has a mixture of two distributions on the types of A, the uniform on $[0,1]$ with weight ε and the uniform on $[e_B, 1]$ with weight $(1-\varepsilon)$:[21]

$$\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H \cap \mathbf{e}_B \leq y] = t_A y, \tag{7}$$

$$\boldsymbol{\tau}_B(\theta_B, e_B)[\mathbf{t}_A \leq x] = \begin{cases} 1, & \text{if } x = e_B = 1, \\ (1-\varepsilon)\,\frac{x-e_B}{1-e_B} + \varepsilon x, & \text{if } e_B \leq x \leq 1,\ e_B < 1, \\ \varepsilon x, & \text{if } 0 \leq x < e_B. \end{cases} \tag{8}$$

This type structure does not fully satisfy the requirements of Section 4.2.1; in particular, $\mathbb{E}_{e_B}[\mathbf{t}_A | \mathbf{t}_A > x]$ is weakly (instead of strictly) increasing in e_B. This implies that we have more equilibria than the ones described in Section 4.2.1; however, we can show that in every non-trivial equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ the decision function $\boldsymbol{\sigma}_A$ is increasing and $\boldsymbol{\sigma}_B(\theta^H, \cdot) : [0,1] \to \{S, K\}$ is equivalent to a weakly increasing function, where we say that two decision functions $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ and $\bar{\boldsymbol{\sigma}}_B(\theta^H, \cdot)$ are equivalent if they induce the same endogenous first-order belief for each type t_A, and hence the same endogenous second-order belief for each epistemic type e_B.[22]

Proposition 3 *In the model given by (7)-(8) a non-trivial equilibrium exists iff $\theta^H \geq \frac{4}{3}$. In every non-trivial Bayesian equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ of the model with $\theta^H \geq \frac{4}{3}$, $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$ for every e_B, $\boldsymbol{\sigma}_A$ is increasing and $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ is either weakly increasing or equivalent to a weakly increasing function, where the thresholds $\hat{e}_A$ and $\hat{e}_B^H$ respectively characterizing $\boldsymbol{\sigma}_A$ and $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ are as follows:*

1. *for $\theta^H \in \left[\frac{4}{3}, 2\right]$: $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{1}{2}, 0\right)$, or $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{\theta^H}{4-\theta^H}, \frac{3}{2} - \frac{2}{\theta^H}\right)$;*

2. *for $\theta^H > 2$: $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{1}{2}, 0\right)$.*

Furthermore,

(a) *for every type t_A, $\alpha_A(t_A) = t_A\left(1 - \hat{e}_B^H\right)$;*

(b) *for every epistemic type e_B,*

$$\begin{aligned} \boldsymbol{\beta}_B^{\varnothing}(e_B) &= \left(1 - \hat{e}_B^H\right)\frac{1 + (1-\varepsilon)\,e_B}{2}, \\ \boldsymbol{\beta}_B^{I}(e_B) &= \begin{cases} \left(1 - \hat{e}_B^H\right)\frac{(1+\hat{e}_A)}{2}, & \text{if } e_B \leq \hat{e}_A \leq 1, \\ \left(1 - \hat{e}_B^H\right)\frac{1+(1-\varepsilon)e_B - \varepsilon\hat{e}_A^2}{2(1-\varepsilon\hat{e}_A)}, & \text{if } 0 \leq \hat{e}_A < e_B. \end{cases} \end{aligned}$$

[21] We also assume that $e_B = 0$ has a uniform measure on $[0,1]$ and $e_B = 1$ has a mixture of the uniform measure on $[0,1]$ and the Dirac measure concentrated on 1 with weights ε and $(1-\varepsilon)$. But since each type of A has an atomless marginal belief on $T_B^e = [0,1]$, the beliefs of these extreme types of B are immaterial.

[22] Given (7) and given that in every equilibrium $\boldsymbol{\sigma}_B(\theta^L, e_B) = K$ for each e_B, the decision functions $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ and $\bar{\boldsymbol{\sigma}}_B(\theta^H, \cdot)$ induce the same first-order belief $\boldsymbol{\alpha}_A(t_A)$ for every type t_A if

$$\mu[\{e_B : \boldsymbol{\sigma}_B(\theta^H, e_B) = S\}] = \mu[\{e_B : \bar{\boldsymbol{\sigma}}_B(\theta^H, e_B) = S\}]$$

where μ is the Lebesgue measure on $T_B^e = [0,1]$.

Comments and sketch of proof First of all, we observe that $\boldsymbol{\beta}_B^I$ is only weakly increasing in e_B; in particular, all the epistemic types of B that are smaller than $\hat{e}_A$ hold the same second-order endogenous belief. This is what gives rise to the multiplicity of equilibria that are equivalent to the threshold equilibria described in the proposition: Type (θ^H, e_B) is indifferent iff $\theta^H \boldsymbol{\beta}_B^I(e_B) = 1$. If (θ^H, e_B) is indifferent and $e_B \leq \hat{e}_A$, then every other type (θ^H, e'_B) with $e'_B \leq \hat{e}_A$ is also indifferent because $\boldsymbol{\beta}_B^I(e_B) = \boldsymbol{\beta}_B^I(e'_B)$. The shape of $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ affects equilibrium beliefs only through the measure of the set of indifferent types that choose S. Any change in $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ that does not change this measure is immaterial.

Moreover, note that in this parametric specification there are two (equivalence classes of) equilibria for $\theta^H \in \left[\frac{4}{3}, 2\right]$. For $\theta^H > 2$ the threshold equilibrium characterized by $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{1}{2}, 0\right)$ is the unique non-trivial equilibrium, as stated in Proposition 2. Finally, non-trivial equilibria do not arise in this setting when $\theta^H < \frac{4}{3}$. The parameter thresholds that define the multiple-equilibrium region have a quite intuitive interpretation. Recall that, in every non-trivial equilibrium, type t_A chooses I iff $\boldsymbol{\alpha}_A(t_A) > \frac{1}{2}$; which implies $\boldsymbol{\beta}_B^I(e_B) > \frac{1}{2}$. As discussed in Section 4.2.3, if $\theta^H = 2$ the high-guilt types prefer to cooperate when $\boldsymbol{\beta}_B^I(e_B) > \frac{1}{2}$. Threshold $\theta^H = \frac{4}{3}$, instead, is related to the assumption that the conditional beliefs on $\mathbf{t}_A$ (given $\mathbf{t}_A \geq \hat{e}_A$) held by any epistemic type $e_B \leq \hat{e}_A$ are described by the uniform distribution on $[\hat{e}_A, 1]$. The easiest way to understand this is to focus on the $(\hat{e}_A, \hat{e}_B^H) = \left(\frac{1}{2}, 0\right)$ equilibrium. In such equilibrium A-types coincide with the first-order endogenous beliefs $(\boldsymbol{\alpha}_A(t_A) = t_A)$, and A chooses I only for types larger than $\frac{1}{2}$; this implies, for every $e_B \leq \hat{e}_A$, $\boldsymbol{\beta}_B^I(e_B) = \frac{3}{4}$, which is the expected value of the uniform distribution on $\left[\frac{1}{2}, 1\right]$. Therefore the lowest guilt sensitivity that makes B willing to Share when his epistemic type is $e_B \leq \hat{e}_A$ is $\theta_B = \frac{4}{3}$. If $\theta^H < \frac{4}{3}$, type $(\theta_B, e_B) = (\theta^H, \hat{e}_A)$ prefers to Keep, and the same must hold for every lower epistemic type. This, together with the observation that there are no equilibria in which the type who is indifferent between S and K is higher than $\hat{e}_A$, allows us to conclude that all the equilibria are trivial when $\theta^H < \frac{4}{3}$. A more formal proof is contained in Appendix A.3.

5 Role-independent guilt

In Section 4 we analyzed a model in which player A is commonly known to be selfish, and only player B can feel guilt. We interpreted this model as the description of a population where the potential guilt sensitivity can be either high or low, whereas the actual guilt sensitivity is triggered by the role in the game. Now we assume instead that potential and actual guilt sensitivity coincide, because actual guilt sensitivity is role-independent. Each individual is affected by simple guilt aversion with guilt sensitivity $\theta \in \Theta$. If an individual with guilt type θ is assigned to role $i \in \{A, B\}$ then $\theta_i = \theta$. Therefore in this model also player A may experience guilt feelings that are triggered by the expectation of B's disappointment. Player B can only be disappointed after the terminal history O, in which case the extent of his disappointment also depends on his strategy. More precisely, B's disappointment depends on his first-order belief on his own choice, i.e. on what B initially plans to do. Here we assume that B's plan coincides with his actual strategy.

To derive B's disappointment, first note that his expected material payoff is

$$\mathbb{E}_B[\mathbf{m}_B] = \begin{cases} 1 \cdot (1 - \alpha_B) + 2 \cdot \alpha_B, & \text{if } s_B = S, \\ 1 \cdot (1 - \alpha_B) + 4 \cdot \alpha_B, & \text{if } s_B = K. \end{cases}$$

Since $m_B(O) = 1$ is the lowest material payoff for B, $\mathbb{E}_B[\mathbf{m}_B] \geq m_B(O)$ and B's disappointment after O is

$$\max\{0, \mathbb{E}_B[\mathbf{m}_B] - m_B(O)\} = \mathbb{E}_B[\mathbf{m}_B] - 1 = \begin{cases} \alpha_B, & \text{if } s_B = S, \\ 3\alpha_B, & \text{if } s_B = K. \end{cases}$$

We can represent this strategic situation with a psychological game parametrized by the guilt sensitivity parameters θ_i $(i = A, B)$. To analyze such (more general) version of the Trust Minigame with guilt aversion we need to expand our notation about beliefs by introducing a feature of Ann's second-order beliefs that describes her expectation of Bob's disappointment if she goes Out:

$$\bar{\beta}_A = \mathbb{E}_A \left[\mathbb{E}_B[\mathbf{m}_B] - m_B(O) \right].$$

The psychological game with role-independent guilt-aversion is more easily represented in a sort of reduced form where each player's psychological utility depends on his own endogenous second-order belief rather than the co-player's endogenous first-order belief, as shown in the Figure 3.

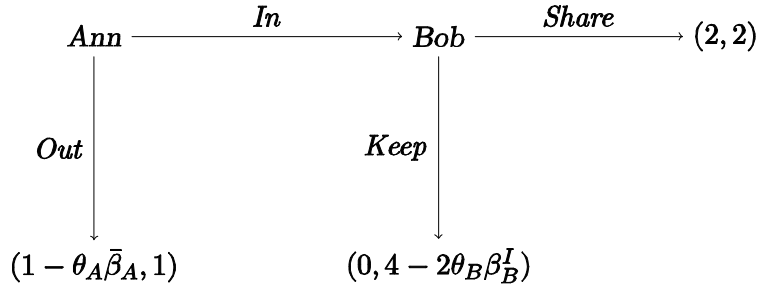

Figure 3. The Trust Minigame with psychological utilities of A and B

5.1 Complete information model

We begin our analysis of the role-independent guilt model by considering its complete information benchmark, as we did for the role-dependent guilt case. We assume therefore that $\theta_A > 0$ and $\theta_B > 0$ are commonly known. As we stressed in Section 4.1, in a complete information equilibrium first and second-order beliefs are correct, and players maximize given their belief.

First of all, (O, K, α, β) is an equilibrium if $\alpha_A = \beta_B^{\varnothing} = 0$, $\alpha_B = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$. If $\theta_A \in \left(0, \frac{1}{3}\right)$ and $\theta_B \in (0, 1)$ this is the unique equilibrium.

If $\theta_B \in [1, +\infty)$, as in the role-dependent guilt case, there is another pure strategy equilibrium (I, S, α, β) where $\alpha_A = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = 1$ and $\bar{\beta}_A = 1$.

Finally, we have one additional pure strategy equilibrium if $\theta_A \in \left[\frac{1}{3}, +\infty\right)$, given by (I, K, α, β) where $\alpha_A = \beta_B^{\varnothing} = 0$, $\alpha_B = 1$, $\bar{\beta}_A = 3$, and $\beta_B^I < \frac{1}{\theta_B}$.

In Appendix B.1 we provide a complete characterization of the equilibrium correspondence, including mixed equilibria.

5.2 Incomplete information model

5.2.1 Type structure

In the analysis of the game with role-independent guilt we maintain several assumptions that we made in Section 4.2.1. In particular, we still have a continuum of types on both sides, and each player i is characterized by a guilt type $\theta_i \in \{\theta^L, \theta^H\}$, with $\theta^L = 0 < \theta^H$, and an epistemic type e_i, where the epistemic type determines the beliefs of player i about the type of the co-player. We assume for simplicity that each type of each player i believes that the guilt and epistemic types of the co-player $-i$ are independent.[23] Specifically, we let e_i parametrize the subjective probability of the high-guilt type of the co-player: $e_i = \mathbb{P}_{(\theta_i, e_i)} \left[\boldsymbol{\vartheta}_{-i} = \theta^H \right]$. This implies that for each i, $t_i = (\theta_i, e_i) \in \{\theta^L, \theta^H\} \times [0, 1] = T_i$ and we can write $\boldsymbol{\tau}_i : [0, 1] \to \Delta\left(\{\theta^L, \theta^H\} \times [0, 1]\right)$.

[23] Thus, we extend to both players the assumption that we made on A's beliefs in the model with role-dependent guilt.

As a consequence, the second-order endogenous beliefs of players A and B are independent of their guilt sensitivity. We also assume that each type of each player has the same marginal beliefs about the epistemic type of the co-player given by a *continuous* cdf F with *full support*. Thus,

$$\forall e_i \in [0,1], \forall x,\, \boldsymbol{\tau}_i(e_i)[\boldsymbol{\vartheta}_{-i} = \theta^H \cap \mathbf{e}_{-i} \leq x] = e_i F(x). \tag{9}$$

As in the previous Section, higher epistemic types have higher beliefs about the type of the co-player. But unlike the previous Section, all the types of player B have the same marginal beliefs about the epistemic type of player A. Here the epistemic type of B parametrizes a different feature of B's beliefs, i.e. the subjective probability that the guilt type of A is high.

By eq. (9), i's expectation of $\mathbf{e}_{-i}$ is independent of e_i, hence we write $\mathbb{E}_{e_i}[\mathbf{e}_{-i}] = \mathbb{E}[\mathbf{e}_{-i}]$. To simplify the exposition and avoid tedious discussions of subcases in the equilibrium analysis, we assume that this expectation is not too low:

$$\mathbb{E}[\mathbf{e}_{-i}] > \frac{1}{3}. \tag{10}$$

5.2.2 Equilibrium analysis

Once again, as explained in Section 2, there is always a *pooling equilibrium with no trust and no cooperation:* all A-types go Out and all B-types Keep. Indeed, if player A is certain that B expects O and that, if surprised by I, he would Keep, O is the material-payoff maximizing choice and A feels no guilt in going Out because of the belief that B is not disappointed. In turn, B is certain that A expects him to Keep (and he may feel certain of this also after observing I with no violation of Bayes rule), hence he feels no guilt for keeping the money because he thinks he is not disappointing A.

Next we study the non-trivial equilibria. Recall that a non-trivial equilibrium is given by a pair of decision functions $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$, such that $\boldsymbol{\sigma}_A^{-1}(I)$ has positive measure; this in turn determines all the endogenous belief functions. In particular we focus on $\boldsymbol{\alpha}_A$, $\boldsymbol{\alpha}_B$, $\bar{\boldsymbol{\beta}}_A$, $\boldsymbol{\beta}_B^{\varnothing}$, $\boldsymbol{\beta}_B^I$; in equilibrium $\boldsymbol{\sigma}_A(\theta_A, e_A)$ is a best reply to $\boldsymbol{\alpha}_A(e_A)$ and $\bar{\boldsymbol{\beta}}_A(e_A)$ for all θ_A and e_A, and $\boldsymbol{\sigma}_B(\theta_B, e_B)$ is a best reply to $\boldsymbol{\beta}_B^I(e_B)$ for all θ_B and e_B.

We show that in this model all non-trivial equilibria are monotone, with $\boldsymbol{\sigma}_A(\cdot,\cdot)$ increasing in both arguments, $\boldsymbol{\sigma}_B(\theta^L,\cdot) = K$ and $\boldsymbol{\sigma}_B(\theta^H,\cdot)$ *decreasing*, because $\boldsymbol{\beta}_B^I$ is decreasing. The reason for the latter is that in this model (unlike the role-dependent guilt model) e_B parametrizes the subjective probability that the guilt type of A is high. A high-guilt type of A is more willing to move In compared to a low-guilt type because this move does not disappoint B. Hence high-guilt types with intermediate first-order beliefs move In, while low-guilt types with the same beliefs stay Out. This implies that the surer B is that A's guilt type is high, the more he thinks that A moves In to avoid guilt rather than to obtain a high material payoff.

The fact that non-optimistic high-guilt types move In to avoid guilt also implies that a non-trivial equilibrium exists for a wide range of parameter values.

Proposition 4 *In the model given by (9)-(10) there is a non-trivial equilibrium iff* $\theta^H > \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$.[24] *If* $\theta^H \in \left(\frac{1}{3\mathbb{E}[\mathbf{e}_B]}, 1\right]$ *there exists a unique non-trivial equilibrium in which* $\boldsymbol{\sigma}_A\left(\theta^L, e_A\right) = O$, $\boldsymbol{\sigma}_A\left(\theta^H, e_A\right) = I$, *and* $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = \boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = K$ *for every* e_A *and* e_B, *which yields the following endogenous beliefs:* $\boldsymbol{\alpha}_A(e_A) = 0$, $\bar{\boldsymbol{\beta}}_A(e_A) = 3\mathbb{E}[\mathbf{e}_B]$, $\boldsymbol{\alpha}_B(e_B) = e_B$ *and* $\boldsymbol{\beta}_B^{\varnothing}(e_B) = \boldsymbol{\beta}_B^I(e_B) = 0$ *for every* e_A *and* e_B. *If* $\theta^H > 1$ *there may be other non-trivial equilibria* $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$; *they all have the following structure:* $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$ *for every* e_B, *the decision functions* $\boldsymbol{\sigma}_A(\theta^k,\cdot)$ *(*$k \in \{L,H\}$*) are weakly increasing,* $\boldsymbol{\sigma}_B\left(\theta^H,\cdot\right)$ *is weakly decreasing, and they are respectively characterized by thresholds* $\hat{e}_A^L$, $\hat{e}_A^H$ *and* $\hat{e}_B^H$, *with* $0 \leq \hat{e}_A^H < \hat{e}_A^L \leq 1$ *and* $0 \leq \hat{e}_B^H < 1$, *such that*

[24] All expectations not indexed by the epistemic type e_i are determined by the common marginal cdf F on $T_{-i}^e = [0,1]$.

(a) *for every epistemic type* e_A,

$$\boldsymbol{\alpha}_A(e_A) = e_A F\left(\hat{e}_B^H\right),$$

$$\begin{aligned}\bar{\boldsymbol{\beta}}_A(e_A) &= \left(1 - F\left(\hat{e}_A^L\right)\right)\left(3 - 2F\left(\hat{e}_B^H\right)e_A\right) + \left(F\left(\hat{e}_A^L\right) - F\left(\hat{e}_A^H\right)\right)3\mathbb{E}\left[\mathbf{e}_B\right] \\ &\quad -2\left(F\left(\hat{e}_A^L\right) - F\left(\hat{e}_A^H\right)\right)e_A F\left(\hat{e}_B^H\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right],\end{aligned}$$

hence $\boldsymbol{\alpha}_A(\cdot)$ *is increasing,* $\bar{\boldsymbol{\beta}}_A(\cdot)$ *is decreasing and the incentive conditions for* A *yield*

$$\hat{e}_A^L = \min\left\{1, \frac{1}{2F\left(\hat{e}_B^H\right)}\right\},$$

$$\hat{e}_A^H > 0 \Rightarrow 1 - \theta^H \bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) = 2\boldsymbol{\alpha}_A\left(\hat{e}_A^H\right);$$

(b) *for every epistemic type* e_B,

$$\boldsymbol{\alpha}_B(e_B) = \left(1 - F\left(\hat{e}_A^L\right)\right) + \left(F\left(\hat{e}_A^L\right) - F\left(\hat{e}_A^H\right)\right)e_B,$$

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = F(\hat{e}_B^H)\mathbb{E}[\mathbf{e}_A],$$

$$\boldsymbol{\beta}_B^I(e_B) = F(\hat{e}_B^H)\left(\mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right]e_B + \mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right](1 - e_B)\right),$$

hence $\boldsymbol{\alpha}_B(\cdot)$ *is increasing,* $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ *is constant,* $\boldsymbol{\beta}_B^I(\cdot)$ *is decreasing, and the incentive condition for* B *yields*

$$0 < \hat{e}_B^H < 1 \Rightarrow \boldsymbol{\beta}_B^I(\hat{e}_B^H) = \frac{1}{\theta^H}.$$

Sketch of proof First note that a low-guilt type of B always Keeps: $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$ for every e_B. This implies that in a non-trivial equilibrium the first-order endogenous belief of A, $\boldsymbol{\alpha}_A(e_A)$, is increasing, as it was in the role-dependent guilt model. Eq. (9) implies that, for every player i, choice c and guilt type θ, the probability of $\boldsymbol{\sigma}_i = c$ given $\boldsymbol{\vartheta}_i = \theta$ is determined by the marginal cdf F; hence we write $\mathbb{P}_F\left[\boldsymbol{\sigma}_i = c|\boldsymbol{\vartheta}_i = \theta\right]$. By eq. (9), the first-order belief of e_A is

$$\boldsymbol{\alpha}_A(e_A) = \mathbb{P}_F\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H\right]e_A.$$

With this, the decision function of the low-guilt type of A is increasing: $\boldsymbol{\sigma}_A(\theta^L, e_A) = I$ if and only if $\boldsymbol{\alpha}_A(e_A) > \frac{1}{2}$, that is

$$e_A > \frac{1}{2\mathbb{P}_F\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H\right]}.$$

Moreover, the comparison between A's incentive condition when her guilt type is low, and A's incentive condition when her guilt type is high implies that if I is optimal for type (θ^L, e_A) then it is also optimal for type (θ^H, e_A). Therefore

$$\mathbb{P}_F\left[\boldsymbol{\sigma}_A = I|\boldsymbol{\vartheta}_A = \theta^H\right] > \mathbb{P}_F\left[\boldsymbol{\sigma}_A = I|\boldsymbol{\vartheta}_A = \theta^L\right].$$

This implies that $\boldsymbol{\alpha}_B$ is increasing: the higher is the probability that B assigns to $\boldsymbol{\vartheta}_A = \theta^H$, the higher is B's belief that A chooses I, because A chooses I for a set of epistemic types that has a larger measure when the guilt type is high. Furthermore, $\boldsymbol{\beta}_B^I$ and $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ must be decreasing: as explained above the higher is e_B, the more B explains choice I with A's desire to avoid guilt rather than obtaining a high material payoff from B's Sharing. Therefore higher values of e_B are associated to lower conditional beliefs on $\boldsymbol{\alpha}_A$, that is to lower $\boldsymbol{\beta}_B^I$.

With this, we can show that $\bar{\boldsymbol{\beta}}_A$ is decreasing and $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is weakly increasing. In particular, we know that $\boldsymbol{\sigma}_A(\theta^L, \cdot)$ is weakly increasing, and that $\{e_A : \boldsymbol{\sigma}_A(\theta^L, e_A) = I\} \subset \{e_A : \boldsymbol{\sigma}_A(\theta^H, e_A) = I\}$. We then conclude that either $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is increasing or $\boldsymbol{\sigma}_A(\theta^H, e_A) = I$ for

every e_A. $\bar{\boldsymbol{\beta}}_A$ is decreasing because the more A believes that B's guilt type is high (the higher e_A), the more he believes that B plans to Share, and – other things being equal – the types that plan to Share have a lower expected material payoff than the types who plan to Keep, therefore they are also less disappointed if A goes Out.

Finally we show that all equilibria are trivial if $\theta^H \leq \frac{1}{3\mathbb{E}[e_B]}$. By eq. (10) $\frac{1}{3\mathbb{E}[e_B]} < 1$. We already know that if $\theta^H < 1$ all B-types Keep. Therefore, for this range of parameters, $\boldsymbol{\alpha}_A = 0$ and A would go In only to avoid guilt, i.e., only a high-guilt type of A may choose I. Formally $\hat{e}_B^H = 0$ and $\hat{e}_A^L = 1$ (recall that $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ is weakly decreasing and $\boldsymbol{\sigma}_A(\theta^L, \cdot)$ is weakly increasing). The fraction of high-guilt types of A going In is positive if $1 - \theta^H \bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) < 0$, where $\bar{\boldsymbol{\beta}}_A(\hat{e}_A^H)$ is the disappointment of B expected by threshold type $(\theta^H, \hat{e}_A^H)$ if he chooses Out. When all B-types Keep, the disappointment of B if A goes Out is

$$(1 \cdot (1 - \boldsymbol{\alpha}_B) + 4 \cdot \boldsymbol{\alpha}_B) - 1 = 3\boldsymbol{\alpha}_B,$$

and the expression for $\bar{\boldsymbol{\beta}}_A(\hat{e}_A^H)$ can be simplified to

$$\bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) = \mathbb{E}_{\hat{e}_A^H}[3\boldsymbol{\alpha}_B] = 3\mathbb{P}_F\left[\boldsymbol{\sigma}_A = I | \boldsymbol{\vartheta}_A = \theta^H\right] \mathbb{E}[\mathbf{e}_B] \leq 3\mathbb{E}[\mathbf{e}_B].$$

Therefore condition $1 - \theta^H \bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) < 0$ can be satisfied only if $\theta^H > \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$.

A more formal proof is contained in Appendix B.2.

6 Discussion

In this paper we analyzed Bayesian equilibrium models of the Trust Minigame with guilt aversion, assuming that each player is uncertain about the guilt sensitivity of the co-player (first-order uncertainty) and/or about the co-player's beliefs about his own guilt sensitivity (second-order uncertainty). The beliefs ascribed to players are determined by their Harsanyi types and are subjective. We analyzed two models: In the first one guilt is *role-dependent*, because only player B, the second mover (or trustee), can feel guilt. In this model A's type is purely epistemic and parametrizes his subjective probability that the guilt type of B is high. On the other hand, B's type has both a guilt and an epistemic component, where the latter parametrizes his second-order exogenous belief, i.e. his belief about A's belief about B's guilt type. In the second model guilt is *role-independent*, i.e. also the first mover A can feel guilt if he thinks that he is disappointing B by not trusting him. For the sake of simplicity and to keep symmetry between A and B also in modeling exogenous beliefs, in this second model we let the epistemic component of both players' types parametrize their first-order exogenous belief, i.e. their subjective probability that the co-player's guilt type is high. Therefore, the second model is not a generalization of the first one.

In the rest of this Section we first discuss the empirical implications of our models and then we offer our methodological perspective on the use of the subjective Bayesian equilibrium concept.

6.1 Empirical predictions

An equilibrium specifies actions, beliefs about actions (endogenous first-order beliefs) and beliefs about beliefs about actions (endogenous second-order beliefs) for each type of each player. We focused on non-trivial equilibria of the Trust Minigame where a positive fraction of A-types trust the second mover, B. Qualitative predictions about behavior and hierarchical beliefs about behavior can be obtained assuming that the actual distribution of types satisfies some mild assumptions. Such predictions can be used to organize experimental data. If the distribution of types has a rich support and the upper bound on guilt aversion is sufficiently high, we should expect not only heterogeneous behavior, but also heterogeneous hierarchical beliefs about behavior,

with a lot of subjects who exhibit intermediate beliefs. Furthermore, if the epistemic component of players' types is statistically independent of the guilt component, then we should observe positive correlation between pro-social actions and endogenous second-order beliefs (cf. Charness & Dufwenberg 2006). Indeed, the willingness to choose the pro-social action, in particular the willingness to Share of B, is an increasing function of the guilt type and of the endogenous (conditional) second-order belief. In our model, the latter depends only on the epistemic type. If epistemic and guilt types are statistically independent, then the pro-social action must be positively correlated with the endogenous second-order belief.[25]

Statistical independence between the guilt and epistemic component of types is a natural benchmark. But it is also plausible to assume that, by a kind of false consensus effect (see Ross *et al.* 1977), types with higher guilt aversion tend to have higher beliefs about the aversion to guilt of the co-player. With such positive correlation, our model with role-independent guilt shows that the endogenous conditional second-order beliefs of B-subjects may be negatively correlated with their guilt type: high-guilt types of B tend to believe that the guilt type of A is high and to explain A's trust as a desire to not disappoint B rather than to obtain a higher material payoff. This tends to decrease the correlation between the pro-social action and (conditional) second-order beliefs. On the other hand, in the model with role-dependent guilt lack of independence may be due to a different kind of false consensus: the higher the guilt type of B, the higher (in the stochastic sense) his belief about A's belief that B's guilt type is high. In this case positive correlation between the guilt and epistemic component tends to strengthen the positive correlation between the pro-social action and endogenous conditional second-order belief.

The actual existence of a false consensus effect does not imply that players' subjective beliefs must display a *perception* of false consensus for the co-player. Such perceptions are modeled by the type structure. In our models there is no perception of false consensus because of the twin assumptions that the belief maps do not depend on the guilt component of players' type, and that each player deems the epistemic component of the co-player type to be independent of the guilt component. Taking into account what we just said about the actual false consensus effect, we can speculate about the effect of introducing the perception of false consensus in our models. If in the model with role-dependent guilt we let A perceive a positive correlation between the guilt and epistemic components of B's type, the qualitative results do not change: now A expects high-guilt types of B to be even more cooperative because he expects them to hold on average higher endogenous second-order beliefs. On the other hand, the effects of introducing a strong perception of false consensus in the model with role-independent guilt are not clear: here higher guilt types of B should be expected to hold on average lower endogenous second-order beliefs.

6.2 Adequacy of subjective Bayesian equilibrium

Our use of Bayesian equilibrium analysis to model behavior and endogenous beliefs deserves discussion. It is sometimes argued that agents learn equilibrium behavior by playing a game many times against randomly matched co-players. However, our analysis cannot rely on such arguments for several reasons. First, in so far as we aim at organizing experimental data, we must take into account that in most experiments on the Trust Game subjects play the game one shot, hence they cannot learn. Second, as noted by Battigalli & Dufwenberg (2009), once behavior has stabilized in a recurrent game, strategy distributions should look like a self-confirming equilibrium, which is likely to be different from a Nash or Bayesian equilibrium if agents have belief-dependent preferences. A third, related issue is that we assume that players do not know the objective distribution of types. Then, even with standard preferences, subjective Bayesian equilibrium is not the right tool to capture selfconfirming patterns of behavior. The

[25] Formally, $\mathbb{P}[\boldsymbol{\sigma}_B = S] = f(\beta, \theta)$, where f is increasing in both arguments. If $\beta = \boldsymbol{\beta}(e)$ and the random variables $\mathbf{e}$ and $\boldsymbol{\vartheta}$ are independent, then $\boldsymbol{\sigma}_B$ must be positively correlated with $\boldsymbol{\beta}$.

reason is that Bayesian equilibrium postulates that players have correct conjectures about the true (type-dependent) decision functions of co-players. This assumption is justified by learning in those (rare) circumstances when agents obtain sufficient information feedback to identify such decision functions.. However, such fine information feedback should also allow to identify the distribution of types (cf. Dekel *et al.* 2004).

We use Bayesian equilibrium analysis to provide an orderly and consistent description of strategic reasoning in an incomplete information environment. It has been shown that, if one drops the assumption that exogenous beliefs are derived from an objective distribution – as we do – then the Bayesian equilibrium assumption that players hold correct conjectures about the co-players' decision functions just ensures that behavior and endogenous beliefs are consistent with common certainty of rationality, which is characterized by incomplete-information rationalizability (Brandenburger & Dekel 1987, Battigalli & Siniscalchi 2003). Of course, our specific assumptions about exogenous beliefs yield equilibrium implications that go beyond mere rationalizability. Therefore we offer an analysis in between objective Bayesian-Nash equilibrium and the most general notion of incomplete-information rationalizability. It would be interesting to explore a rationalizability approach to the Trust Minigame with guilt aversion whereby some restrictions on beliefs are taken as given and commonly understood, as suggested by Battigalli & Siniscalchi (2003) for games with standard preferences. Battigalli *et al.* (2012) essentially is an example of this approach to the analysis of a cheap-talk sender-receiver game where the sender is affected by guilt aversion.

Appendix

Appendix A. Analysis of role-dependent guilt

A.1: Equilibrium with complete information

We describe the mixed equilibrium correspondence of the Trust Minigame with guilt aversion when (θ_A, θ_B) is common knowledge, $\theta_A = 0$ and $\theta_B > 0$. We rely on Nash's **mass-action** interpretation (cf. Weibull 1996) and think of a mixed strategy $\sigma_i \in \Delta(S_i)$ as coming from a statistical distribution of pure strategies in a population of individuals playing in role i, under the assumption that individuals are drawn at random and matched to play the game. Thus $\sigma_i(s_i)$ is the fraction of individuals in population i playing s_i, and also the objective probability that s_i is played; but no individual actually randomizes. Thus, for example, if player A carries out his plan and O occurs, A cannot be disappointed because this means that the individual playing in role A planned to choose O with probability one. This interpretation is consistent with the incomplete information analysis to follow. The main difference between the equilibria analyzed here and those of the incomplete information models is that here all the individuals playing in role i have the same beliefs about the co-player.

Given the special form of our psychological utility functions, we would obtain the same equilibria under the assumption that $\sigma_i(s_i)$ is the probability with which i plans to choose s_i. Such equivalence holds trivially for standard games, but it does not hold for all psychological games.

An **equilibrium** is a profile $(\sigma_A, \sigma_B, \alpha_A, \alpha_B, \beta_B^{\varnothing}, \beta_B^I)$ that satisfies the incentive conditions

$$\begin{aligned}
\sigma_A(I) &> 0 \Rightarrow \alpha_A \geq \frac{1}{2}, \\
\sigma_A(I) &< 1 \Rightarrow \alpha_A \leq \frac{1}{2}, \\
\sigma_B(S) &> 0 \Rightarrow \beta_B^I \geq \frac{1}{\theta_B}, \\
\sigma_B(S) &< 1 \Rightarrow \beta_B^I \leq \frac{1}{\theta_B},
\end{aligned}$$

and the belief conditions

$$\begin{aligned}
\alpha_A &= \sigma_B(S), \\
\alpha_B &= \sigma_A(I), \\
\beta_B^{\varnothing} &= \alpha_A, \\
\alpha_B &> 0 \Rightarrow \beta_B^I = \beta_B^{\varnothing}.
\end{aligned}$$

Under the mass-action interpretation, the incentive conditions say that an action can be chosen by a positive fraction of individuals in population i only if it is a best reply to the common belief about $-i$. The belief conditions state that beliefs of the first and second-order are correct. In particular, the overall initial second-order belief of B is a joint probability measure on the strategies and first-order beliefs of A, say $\mu_B^2 \in \Delta(\{I, O\} \times [0,1])$, where $[0,1]$ is the set of possible values of α_A and hence represents the set of first-order beliefs of A. In a complete-information equilibrium with first-order beliefs (α_A, α_B), μ_B^2 assigns marginal probability α_B to I and marginal probability 1 to the true first-order belief α_A, that is

$$\mu_B^2(I \times [x,y]) = \begin{cases} \alpha_B, & \text{if } \alpha_A \in [x,y], \\ 0, & \text{if } \alpha_A \notin [x,y], \end{cases}$$

for each interval $[x,y] \subseteq [0,1]$. This implies $\beta_B^{\varnothing} = \alpha_A$ and $\beta_B^I = \alpha_A = \beta_B^{\varnothing}$ if $\alpha_A = \mu_B^2(I \times [0,1]) > 0$. The latter condition holds because if $\mu_B^2(I \times [0,1]) > 0$ then $\mu_B^2(\cdot|I)$ must be a Dirac measure supported by α_A:

$$\mu_B^2([x,y]|I) = \frac{\mu_B^2(I \times [x,y])}{\mu_B^2(I \times [0,1])} = \begin{cases} 1, & \text{if } \alpha_A \in [x,y], \\ 0, & \text{if } \alpha_A \notin [x,y]. \end{cases}$$

With this, the mixed equilibrium correspondence is as follows:

- If $\theta_B < 1$, then $\alpha_B = \sigma_A(I) = 0$, $\beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = 0$, and β_B^I is arbitrary.
- If $1 \leq \theta_B < 2$, then
  - either $\alpha_B = \sigma_A(I) = 0$, $\beta_B^{\varnothing} = \alpha_A = \sigma_B(S) \leq \frac{1}{2}$, $\beta_B^I \leq \frac{1}{\theta_B}$, and $\sigma_B(S) > 0 \Rightarrow \beta_B^I = \frac{1}{\theta_B}$;
  - or $\alpha_B = \sigma_A(I) = 1$, $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = \frac{1}{\theta_B}$;
  - or $\alpha_B = \sigma_A(I) = 1$, $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = 1$.
- If $\theta_B = 2$, then
  - either $\alpha_B = \sigma_A(I) = 0$, $\beta_B^{\varnothing} = \alpha_A = \sigma_B(S) \leq \frac{1}{2}$, $\beta_B^I \leq \frac{1}{2}$, and $\sigma_B(S) > 0 \Rightarrow \beta_B^I = \frac{1}{2}$;
  - or $0 < \alpha_B = \sigma_A(I) < 1$, $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = \frac{1}{2}$;
  - or $\alpha_B = \sigma_A(I) = 1$, $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = 1$.

- If $\theta_B > 2$, then
  - either $\sigma_A(I) = \alpha_B = 0$, $\beta_B^{\varnothing} = \alpha_A = \sigma_B(S) \leq \frac{1}{2}$, $\beta_B^I \leq \frac{1}{\theta_B}$, and $\sigma_B(S) > 0 \Rightarrow \beta_B^I = \frac{1}{\theta_B}$,
  - or $\alpha_B = \sigma_A(I) = 1$, $\beta_B^I = \beta_B^{\varnothing} = \alpha_A = \sigma_B(S) = 1$.

A.2: Proof of Proposition 1

We start from the conjecture that set $\{t_A \in [0,1] : \boldsymbol{\sigma}_A(t_A) = I\}$ has a strictly positive measure and provide a characterization of the equilibria that verify this property. We go through a sequence of claims.

Claim 5 *For every e_B, $\boldsymbol{\sigma}_B(\theta^L, e_B) = K$, and $\boldsymbol{\sigma}_B(\theta^H, e_B) = K$ whenever $\theta^H \leq 1$.*

Proof Fix e_B arbitrarily. Recall that the optimal choice of type (θ_B, e_B) is K if $\theta_B \boldsymbol{\beta}_B^I(e_B) < 1$. Therefore $\boldsymbol{\sigma}_B(\theta^L, e_B) = K$, because $\theta^L < 1$ and $\boldsymbol{\beta}_B^I(e_B) \in [0,1]$. Next we prove that $\boldsymbol{\beta}_B^I(e_B) < 1$, which implies that $\boldsymbol{\sigma}_B(\theta^H, e_B) = K$ whenever $\theta^H \leq 1$.[26] By assumption, event $[\boldsymbol{\sigma}_A = I] = \{t_A \in [0,1] : \boldsymbol{\sigma}_A(t_A) = I\}$ has strictly positive measure. Therefore $\boldsymbol{\alpha}_B(e_B) = \boldsymbol{\tau}_B(e_B)[\boldsymbol{\sigma}_A = I] > 0$, because cdf F_{e_B} is strictly increasing on $[0,1]$, hence the corresponding measure $\boldsymbol{\tau}_B(e_B) \in \Delta([0,1])$ has full support. This implies that $\boldsymbol{\beta}_B^I(e_B)$ is determined by Bayes rule:

$$\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \boldsymbol{\sigma}_A = I] = \int_{\{t_A \in [0,1] : \boldsymbol{\sigma}_A(t_A) = I\}} \boldsymbol{\alpha}_A(t_A) \boldsymbol{\tau}_B(e_B)[\mathrm{d}t_A | \boldsymbol{\sigma}_A = I],$$

where $\boldsymbol{\tau}_B(e_B)[\cdot | \boldsymbol{\sigma}_A = I]$ is the conditional measure given by

$$\boldsymbol{\tau}_B(e_B)[E | \boldsymbol{\sigma}_A = I] = \frac{\boldsymbol{\tau}_B(e_B)[E \cap (\boldsymbol{\sigma}_A = I)]}{\boldsymbol{\tau}_B(e_B)[\boldsymbol{\sigma}_A = I]},$$

for every measurable set $E \subseteq [0,1]$.

Since $\boldsymbol{\sigma}_B(\theta^L, e'_B) = K$ for every $e'_B \in [0,1]$, every type t_A assigns at least probability $1 - t_A$ to K:

$$1 - \boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B = K] \geq \boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^L] = 1 - t_A.$$

This implies

$$\{t_A \in [0,1] : \boldsymbol{\alpha}_A(t_A) = 1\} \subseteq \{1\}.$$

Since cdf F_{e_B} is continuous, $\boldsymbol{\tau}_B(e_B)$ is an atomless probability measure, hence

$$\boldsymbol{\tau}_B(e_B)[\boldsymbol{\alpha}_A = 1] \leq \boldsymbol{\tau}_B(e_B)[\{1\}] = 0.$$

Therefore

$$\boldsymbol{\tau}_B(e_B)[\boldsymbol{\alpha}_A < 1 \cap \boldsymbol{\sigma}_A = I] = \boldsymbol{\tau}_B(e_B)[\boldsymbol{\sigma}_A = I],$$

which implies

$$\boldsymbol{\beta}_B^I(e_B) = \int_{\{t_A \in [0,1] : \boldsymbol{\sigma}_A(t_A) = I, \boldsymbol{\alpha}_A(t_A) < 1\}} \boldsymbol{\alpha}_A(t_A) \boldsymbol{\tau}_B(e_B)[\mathrm{d}t_A | \boldsymbol{\sigma}_A = I] < 1.$$

□

We now focus on the case in which $\theta^H > 1$ and we analyze the equilibrium functions $\boldsymbol{\sigma}_A(\cdot) : [0,1] \to \{I, O\}$, $\boldsymbol{\sigma}_B(\theta^L, \cdot) : [0,1] \to \{S, K\}$, and $\boldsymbol{\sigma}_B(\theta^H, \cdot) : [0,1] \to \{S, K\}$ to show that they are (weakly) increasing.[27] We also provide a characterization of some properties of the endogenous beliefs of A and B. The following claim shows that $\boldsymbol{\alpha}_A$ and $\boldsymbol{\sigma}_A$ are increasing. We let μ_A denote the probability measure on $T_B^e = [0,1]$ induced by cdf F.[28] Since the random variable

[26] We could use our tie-breaking rule (the low action is chosen when indifferent) to conclude that $\boldsymbol{\sigma}_B(\theta^H, e_B) = K$ even if $\theta^H \boldsymbol{\beta}_B(e_B) = 1$, which yields the desired result. But we prefer the longer proof in the text to show that tie-breaking rules simplify the exposition, but are immaterial for our results.

[27] Recall that we let I and S be the "high" actions of player A and B respectively.

[28] That is, $\mu_A((x,y]) = F(y) - F(x)$ for every $x, y \in T_B^e = [0,1]$ with $x < y$.

$\boldsymbol{\beta}_B^I$ depends only on e_B, to ease notation we write

$$\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] := \mu_A\left[\Theta_B \times \left\{e_B : \boldsymbol{\beta}_B^I(e_B) > \frac{1}{\theta^H}\right\}\right].$$

Claim 6 *The beliefs and decision function of A satisfy the following conditions:*

$$\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] > \frac{1}{2}$$

and

$$\begin{aligned}\boldsymbol{\alpha}_A(t_A) &= t_A\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right],\\ \boldsymbol{\sigma}_A(t_A) &= I \Longleftrightarrow t_A > \frac{1}{2\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right]},\end{aligned}$$

for every t_A.

Proof Claim 5 and the incentive condition for B imply

$$[\boldsymbol{\sigma}_B = S] = \{(\theta_B, e_B) : \theta_B = \theta^H \wedge \theta^H \boldsymbol{\beta}_B^I(e_B) > 1\}.$$

By assumption, $\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H \cap \mathbf{e}_B \leq y] = t_A F(y)$ for each y. Therefore

$$\boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B = S] = \boldsymbol{\tau}_A(t_A)\left[\boldsymbol{\vartheta}_B = \theta^H \cap \boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] = t_A\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right].$$

The incentive condition for A ($\boldsymbol{\sigma}_A(t_A) = I$ iff $\boldsymbol{\alpha}(t_A) > \frac{1}{2}$) yields the equivalence. If $\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] \leq \frac{1}{2}$, then $\boldsymbol{\sigma}_A(t_A) = O$ for every t_A, contradicting the assumption that a positive fraction of A-types choose I. Therefore $\mu_A[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}] > \frac{1}{2}$. □

Claim 7 *Let*

$$\begin{aligned}\hat{e}_A &= \frac{1}{2\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right]},\\ \hat{e}_B^H &= \sup\left\{e_B : \boldsymbol{\beta}_B^I(e_B) \leq \frac{1}{\theta^H}\right\}.\end{aligned}$$

Then, for every e_B

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = \left(1 - F\left(\hat{e}_B^H\right)\right)\int_0^1 t_A \mathrm{d}F_{e_B}(t_A),$$

$$\boldsymbol{\beta}_B^I(e_B) = \frac{1 - F(\hat{e}_B^H)}{1 - F_{e_B}(\hat{e}_A)}\int_{\hat{e}_A}^1 t_A \mathrm{d}F_{e_B}(t_A) > \frac{1}{2},$$

hence $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ and $\boldsymbol{\beta}_B^I(\cdot)$ are strictly increasing.

Proof By Claim 6

$$\begin{aligned}\boldsymbol{\beta}_B^{\varnothing}(e_B) &= \mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right]\int_0^1 t_A \mathrm{d}F_{e_B}(t_A) = \mathbb{E}[\boldsymbol{\alpha}_A],\\ \boldsymbol{\beta}_B^I(e_B) &= \frac{\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right]}{1 - F_{e_B}(\hat{e}_A)}\int_{\hat{e}_A}^1 t_A \mathrm{d}F_{e_B}(t_A) = \mathbb{E}\left[\boldsymbol{\alpha}_A | \boldsymbol{\alpha}_A > \frac{1}{2}\right] > \frac{1}{2},\end{aligned}$$

where the inequality is strict because $\hat{e}_A < 1$, $\boldsymbol{\alpha}_A(\hat{e}_A) = \frac{1}{2}$ and F_{e_B} is strictly increasing. By assumption (6), $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ and $\boldsymbol{\beta}_B^I(\cdot)$ are strictly increasing. By the definition of $\hat{e}_B^H$, $\mu_A\left[\boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] = 1 - F(\hat{e}_B^H)$, which yields the formulas for $\boldsymbol{\beta}_B^{\varnothing}(e_B)$ and $\boldsymbol{\beta}_B^I(e_B)$. □

The rest of the proposition follows from the monotonicity of $\boldsymbol{\beta}_B^I$ and the observation that $\boldsymbol{\sigma}_B(\theta^H, e_B) = S$ iff $\boldsymbol{\beta}_B^I(e_B) > \frac{1}{\theta^H}$. ■

A.3: Proof of Proposition 3

Proof. Remember that $\mathbb{E}_{e_B}[\mathbf{t}_A|\mathbf{t}_A > x]$ is only weakly increasing in e_B; therefore, we cannot use the results of Proposition 1 that depend on (6). Of course, $\boldsymbol{\sigma}_B\left(\theta^L, e_B\right) = K$ for every e_B. Moreover, a statement analogous to Claim 6 holds. The only difference is that here the set of epistemic types of B choosing S when $\boldsymbol{\vartheta}_B = \theta^H$ may be different from the set of e_B such that $\boldsymbol{\beta}_B^I(e_B) > \frac{1}{\theta^H}$. To ease notation, we let

$$E_B^{HS} = \left\{e_B : \boldsymbol{\sigma}_B(\theta^H, e_B) = S\right\}$$

denote the set of B's epistemic types such that Harsanyi type (θ^H, e_B) Shares. From the marginal cdf F we recover the measure of this set according to A's belief, $\mu_A[E_B^{HS}]$.

Claim 8 *Let*

$$\hat{e}_A = \frac{1}{2\mu_A[E_B^{HS}]}.$$

Then, for every e_B

$$\begin{aligned}\boldsymbol{\beta}_B^{\varnothing}(e_B) &= \mu_A[E_B^{HS}]\int_0^1 t\mathrm{d}F_{e_B}(t)\\ &= \mu_A[E_B^{HS}]\frac{1+(1-\varepsilon)e_B}{2},\end{aligned}$$

which is strictly increasing in e_B*, and*

$$\begin{aligned}\boldsymbol{\beta}_B^I(e_B) &= \frac{\mu_A[E_B^{HS}]}{1-F_{e_B}(\hat{e}_A)}\int_{\hat{e}_A}^1 t\mathrm{d}F_{e_B}(t)\\ &= \begin{cases}\mu_A[E_B^{HS}]\frac{(1+\hat{e}_A)}{2}, & \text{if } e_B \le \hat{e}_A \le 1,\\ \mu_A[E_B^{HS}]\frac{1+(1-\varepsilon)e_B-\varepsilon\hat{e}_A^2}{2(1-\varepsilon\hat{e}_A)}, & \text{if } 0 \le \hat{e}_A < e_B,\end{cases}\end{aligned}$$

which is constant for $e_B \le \hat{e}_A$ *and strictly increasing for* $e_B > \hat{e}_A$.

Proof. By (the analog of) Claim 6

$$\begin{aligned}\boldsymbol{\beta}_B^{\varnothing}(e_B) &= \mu_A[E_B^{HS}]\int_0^1 t\mathrm{d}F_{e_B}(t)\\ &= \mu_A[E_B^{HS}]\int_0^1 t\mathrm{d}F_{e_B}(t)\\ &= \mu_A[E_B^{HS}]\left(\int_0^{e_B}\varepsilon t\mathrm{d}t + \int_{e_B}^1\left(\frac{1-\varepsilon}{1-e_B}+\varepsilon\right)t\mathrm{d}t\right)\\ &= \mu_A[E_B^{HS}]\left(\frac{1}{2}\left(\varepsilon e_B^2 + \left(\frac{1-\varepsilon}{1-e_B}+\varepsilon\right)\left(1-e_B^2\right)\right)\right)\\ &= \mu_A[E_B^{HS}]\frac{1+(1-\varepsilon)e_B}{2},\end{aligned}$$

which is strictly increasing in e_B, and

$$\begin{aligned}
\boldsymbol{\beta}_B^I(e_B) &= \frac{\mu_A[E_B^{HS}]}{1-F_{e_B}(\hat{e}_A)}\int_{\hat{e}_A}^1 t\mathrm{d}F_{e_B}(t) \\
&= \begin{cases} \frac{\mu_A[E_B^{HS}]}{\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)(1-\hat{e}_A)}\int_{\hat{e}_A}^1 t\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)\mathrm{d}t, & \text{if } e_B \le \hat{e}_A \le 1, \\ \frac{\mu_A[E_B^{HS}]}{(1-\varepsilon)+\varepsilon(1-\hat{e}_A)}\left(\int_{\hat{e}_A}^{e_B} t\varepsilon\mathrm{d}t + \int_{e_B}^1 t\left(\frac{1-\varepsilon}{1-e_B}+\varepsilon\right)\mathrm{d}t\right), & \text{if } 0 \le \hat{e}_A < e_B \end{cases} \\
&= \begin{cases} \mu_A[E_B^{HS}]\frac{\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)\frac{[t^2]_{\hat{e}_A}^1}{2}}{\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)(1-\hat{e}_A)}, & \text{if } e_B \le \hat{e}_A \le 1, \\ \frac{\mu_A[E_B^{HS}]}{(1-\varepsilon\hat{e}_A)}\left(\varepsilon\frac{[t^2]_{\hat{e}_A}^{e_B}}{2}+\left(\frac{1-\varepsilon}{1-e_B}+\varepsilon\right)\frac{[t^2]_{e_B}^1}{2}\right), & \text{if } 0 \le \hat{e}_A < e_B \end{cases} \\
&= \begin{cases} \frac{\mu_A[E_B^{HS}]}{\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)(1-\hat{e}_A)}\frac{\left(\frac{(1-\varepsilon)}{(1-e_B)}+\varepsilon\right)(1-\hat{e}_A^2)}{2}, & \text{if } e_B \le \hat{e}_A \le 1, \\ \frac{\mu_A[E_B^{HS}]}{(1-\varepsilon\hat{e}_A)}\left(\frac{\varepsilon\left(e_B^2-\hat{e}_A^2\right)+\left(\frac{1-\varepsilon}{1-e_B}+\varepsilon\right)\left(1-e_B^2\right)}{2}\right), & \text{if } 0 \le \hat{e}_A < e_B \end{cases} \\
&= \begin{cases} \mu_A[E_B^{HS}]\frac{(1+\hat{e}_A)}{2}, & \text{if } e_B \le \hat{e}_A \le 1, \\ \mu_A[E_B^{HS}]\frac{1+(1-\varepsilon)e_B-\varepsilon\hat{e}_A^2}{2(1-\varepsilon\hat{e}_A)}, & \text{if } 0 \le \hat{e}_A < e_B. \end{cases}
\end{aligned}$$

which is constant for $e_B \le \hat{e}_A$ and strictly increasing for $e_B > \hat{e}_A$. □

Now we show that in every non-trivial equilibrium, that is an equilibrium with $\hat{e}_A < 1$, every epistemic type of B who is indifferent between S and K (when $\boldsymbol{\vartheta}_B = \theta^H$) is smaller than $\hat{e}_A$.

Claim 9 *If $\hat{e}_A < 1$, for every $\bar{e}_B \in [0,1]$*

$$\theta^H \boldsymbol{\beta}_B^I(\bar{e}_B) = 1 \Longrightarrow \bar{e}_B \le \hat{e}_A.$$

Proof. Suppose that there exists $\bar{e}_B > \hat{e}_A$ such that $\theta^H \boldsymbol{\beta}_B^I(\bar{e}_B) = 1$. By Claim 8, $\boldsymbol{\beta}_B^I$ is strictly increasing on $(\hat{e}_A, 1]$. Therefore, $\theta^H \boldsymbol{\beta}_B^I(e_B) < 1$ for every $e_B < \bar{e}_B$, and $\theta^H \boldsymbol{\beta}_B^I(e_B) > 1$ for every $e_B > \bar{e}_B$. Hence in this case $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ is increasing and characterized by the threshold $\bar{e}_B$. As a consequence we can rewrite A's incentive condition as

$$\hat{e}_A = \frac{1}{2(1-\bar{e}_B)},$$

which does not have a solution in $[0,1]$ if $\bar{e}_B > \hat{e}_A \ge \frac{1}{2}$. □

Let us now check the equilibria that may arise when all the epistemic types of B find optimal to Share when B's guilt type is high (that is when there is no indifferent type).

Claim 10 *There is at most one non-trivial equilibrium in which $\theta^H \boldsymbol{\beta}_B^I(0) > 1$, and therefore $\boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = S$ for each e_B. This is the threshold equilibrium characterized by $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{1}{2}, 0\right)$, and it exists only for $\theta^H > \frac{4}{3}$.*

Proof. Given that all epistemic types of B choose S when $\boldsymbol{\vartheta}_B = \theta^H$, the incentive condition for the threshold type $t_A = \hat{e}_A$ implies $\hat{e}_A = \frac{1}{2}$. Since $\boldsymbol{\sigma}_A$ is increasing and characterized by the threshold $\hat{e}_A = \frac{1}{2}$, a high-guilt type of B always chooses S, given that his second-order endogenous belief is

$$\boldsymbol{\beta}_B^I(e_B) = \begin{cases} \frac{3}{4} > \frac{1}{\theta^H}, & \text{if } e_B \le \frac{1}{2}, \\ \frac{1+(1-\varepsilon)e_B-\frac{\varepsilon}{4}}{2-\varepsilon} > \frac{1}{\theta^H}, & \text{if } e_B > \frac{1}{2}, \end{cases}$$

for $\theta^H > \frac{4}{3}$. □

Claim 11 *Every non-trivial equilibrium in which* $\theta^H \boldsymbol{\beta}_B^I(e_B) = 1$ *for every* $e_B \leq \hat{e}_A$ *is equivalent to a threshold equilibrium with* $\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{\theta^H}{4-\theta^H}, \frac{3}{2} - \frac{2}{\theta^H}\right)$. *Such equilibria exist if and only if* $\theta^H \in \left[\frac{4}{3}, 2\right]$.

Proof. Fix any equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ as in the claim. First recall that by the analog of Claim 6 $\boldsymbol{\sigma}_A$ is increasing with threshold $\hat{e}_A \geq \frac{1}{2}$. Next, note that all epistemic types $e_B \leq \hat{e}_A$ are indifferent between S and K. Since $\boldsymbol{\beta}_B^I$ is weakly increasing (Claim 8), there is an equivalent threshold equilibrium $(\boldsymbol{\sigma}_A, \bar{\boldsymbol{\sigma}}_B)$ where the threshold $\hat{e}_B^H$ characterizing $\bar{\boldsymbol{\sigma}}_B$ satisfies $(1 - \hat{e}_B^H) = \mu_A[E_B^{HS}]$. The equilibrium conditions for $(\boldsymbol{\sigma}_A, \bar{\boldsymbol{\sigma}}_B)$ are:

$$\begin{cases} (1 - \hat{e}_B^H)\frac{(1+\hat{e}_A)}{2} = \frac{1}{\theta^H} \\ 2\hat{e}_A(1 - \hat{e}_B^H) = 1. \end{cases}$$

Solving the indifference conditions above we obtain

$$\left(\hat{e}_A, \hat{e}_B^H\right) = \left(\frac{\theta^H}{4 - \theta^H}, \frac{3}{2} - \frac{2}{\theta^H}\right).$$

Checking for the values of θ^H that satisfy the inequalities $0 \leq \hat{e}_B^H < \hat{e}_A < 1$, we get $\theta^H \in \left[\frac{4}{3}, 2\right]$. □

Claim 12 *Every equilibrium is trivial if* $\theta^H < \frac{4}{3}$.

Proof. By Claim 9 there cannot exist non-trivial equilibria where the indifferent epistemic type is larger than $\hat{e}_A$. By Claims 10 and 11, non-trivial equilibria where all the types $e_B \leq \hat{e}_A$ are indifferent, or where no epistemic type is indifferent, exist only if $\theta^H \geq \frac{4}{3}$. Hence, all the equilibria with $\theta^H < \frac{4}{3}$ are trivial. □

■

Appendix B. Analysis of role-independent guilt

B.1: Equilibria in the complete-information game

We describe the mixed equilibrium correspondence of the Trust Minigame with guilt aversion when (θ_A, θ_B) is common knowledge, $\theta_A > 0$ and $\theta_B > 0$. (We say that guilt aversion is not role-dependent because both A and B can feel guilt and we are not assuming that the guilt parameter is higher for the agent playing in a particular role.) An equilibrium is a profile (σ_A, σ_B) that induces the endogenous beliefs $\left(\alpha_A, \alpha_B, \beta_B^{\varnothing}, \beta_B^I, \bar{\beta}_A\right)$ and that satisfies the incentive conditions

$$\begin{aligned} \sigma_A(I) &> 0 \Rightarrow \alpha_A \geq \frac{1}{2} - \frac{\theta_A \bar{\beta}_A}{2}, \\ \sigma_A(I) &< 1 \Rightarrow \alpha_A \leq \frac{1}{2} - \frac{\theta_A \bar{\beta}_A}{2}, \\ \sigma_B(S) &> 0 \Rightarrow \beta_B^I \geq \frac{1}{\theta_B}, \\ \sigma_B(S) &< 1 \Rightarrow \beta_B^I \leq \frac{1}{\theta_B}, \end{aligned}$$

and the belief conditions

$$\begin{aligned}
\alpha_A &= \sigma_B(S), \\
\alpha_B &= \sigma_A(I), \\
\beta_B^{\varnothing} &= \alpha_A, \\
\alpha_B &> 0 \Rightarrow \beta_B^I = \beta_B^{\varnothing}, \\
\bar{\beta}_A &= \sigma_B(S)\alpha_B + 3\left(1 - \sigma_B(S)\right)\alpha_B.
\end{aligned}$$

In order to have a full analysis of the equilibria, we proceed by considering seven regions in the parameter space.

- If $\theta_B < 1$ and $\theta_A < \frac{1}{3}$, then $\alpha_B = \sigma_A(I) = 0$, $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$.
- If $\theta_B < 1$ and $\theta_A \geq \frac{1}{3}$, then
  - either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = \frac{1}{3\theta_A}$, $\bar{\beta}_A = \frac{1}{\theta_A}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 3$.
- If $1 \leq \theta_B < 2$ and $\theta_A < \frac{1}{3}$, then
  - either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = \frac{3\theta_B - 2}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 1$.
- If $1 \leq \theta_B < 2$ and $\theta_A \geq \frac{1}{3}$, then
  - either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = \frac{1}{3\theta_A}$, $\bar{\beta}_A = \frac{1}{\theta_A}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 3$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = \frac{3\theta_B - 2}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 1$.
- If $\theta_B \geq 2$ and $\theta_A < \frac{\theta_B - 2}{3\theta_B - 2}$, then
  - either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 1$.
- $\theta_B \geq 2$ and $\frac{\theta_B - 2}{3\theta_B - 2} \leq \theta_A < \frac{1}{3}$ then
  - either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;
  - or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = \frac{\theta_B - 2}{\theta_A(3\theta_B - 2)}$, $\bar{\beta}_A = \frac{\theta_B - 2}{\theta_A \theta_B}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = \frac{3\theta_B - 2}{\theta_B}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 1$.

• $\theta_B \geq 2$ and $\theta_A \geq \frac{1}{3}$, then

- either $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = 0$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$, and $\beta_B^I < \frac{1}{\theta_B}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = \frac{1}{3\theta_A}$, $\bar{\beta}_A = \frac{1}{\theta_A}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 0$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 3$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 0$, $\bar{\beta}_A = 0$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = \frac{\theta_B - 2}{\theta_A(3\theta_B - 2)}$, $\bar{\beta}_A = \frac{\theta_B - 2}{\theta_A \theta_B}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = \frac{1}{\theta_B}$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = \frac{3\theta_B - 2}{\theta_B}$;

- or $\alpha_A = \sigma_B(S) = \beta_B^{\varnothing} = \beta_B^I = 1$, $\alpha_B = \sigma_A(I) = 1$, $\bar{\beta}_A = 1$.

B.2: Proof of Proposition 4

We start from the conjecture that a strictly positive fraction of A-types choose I and provide a characterization of the equilibria that verify this property. We analyze the equilibrium functions $\boldsymbol{\sigma}_A(\theta^k, \cdot) : [0,1] \to \{I, O\}$, $\boldsymbol{\sigma}_B(\theta^k, \cdot) : [0,1] \to \{S, K\}$, with $k = H, L$, and we show that they are monotone, with $\boldsymbol{\sigma}_A(\theta^k, \cdot)$ weakly increasing and $\boldsymbol{\sigma}_B(\theta^k, \cdot)$ weakly decreasing (letting I and S be the "high" actions of A and B respectively). We also provide a characterization of some properties of the endogenous beliefs. We do so by proceeding through a series of claims.

First note that the analog of Claim 5 holds. In particular, B Shares iff $\boldsymbol{\vartheta}_B = \theta^H$ and $\boldsymbol{\beta}_B^I \theta^H > 1$. Recall that μ_A is the common marginal belief of each type of A about the epistemic type of B, and $E_B^{HS} = \left\{e_B : \boldsymbol{\sigma}_B(\theta^H, e_B) = S\right\}$.

Claim 13 *For every e_A,*

$$\begin{aligned} \boldsymbol{\alpha}_A(e_A) &= e_A \mu_A[E_B^{HS}], \\ \boldsymbol{\sigma}_A(\theta^L, e_A) &= \begin{cases} I, & \text{if } e_A > \hat{e}_A^L, \\ O, & \text{otherwise,} \end{cases} \end{aligned}$$

where $\hat{e}_A^L = \min\left\{1, \frac{1}{2\mu_A[E_B^{HS}]}\right\} \in \left[\frac{1}{2}, 1\right]$.

Proof The analog of Claim 5 and B's incentive condition imply

$$[\boldsymbol{\sigma}_B = S] = \{(\theta_B, e_B) : \theta_B = \theta^H, \theta^H \boldsymbol{\beta}_B^I(e_B) > 1\}.$$

By assumption, $\boldsymbol{\tau}_A(e_A)[\boldsymbol{\vartheta}_B = \theta^H \cap \mathbf{e}_B \leq y] = e_A F(y)$ for each y. Therefore

$$\boldsymbol{\alpha}_A(e_A) = \boldsymbol{\tau}_A(e_A)[\boldsymbol{\sigma}_B = S] = \boldsymbol{\tau}_A(e_A)\left[\boldsymbol{\vartheta}_B = \theta^H \cap \boldsymbol{\beta}_B^I > \frac{1}{\theta^H}\right] = e_A \mu_A[E_B^{HS}],$$

which is increasing in e_A. The incentive condition for A when the guilt type is low ($\boldsymbol{\sigma}_A(\theta^L, e_A) = I$ iff $\boldsymbol{\alpha}_A(e_A) > \frac{1}{2}$) implies that $\hat{e}_A^L = \min\left\{1, \frac{1}{2\mu_A[E_B^{HS}]}\right\}$; notice that $\hat{e}_A^L \in \left[\frac{1}{2}, 1\right]$.

Next note that in a non-trivial equilibrium A necessarily expects to disappoint B by going Out. Formally:

Claim 14 *For each e_A, $\bar{\boldsymbol{\beta}}_A(e_A) > 0$.*

Proof In a non-trivial equilibrium a positive fraction of A-types go In, i.e., the set

$$\left\{e_A : \boldsymbol{\sigma}_A(\theta^L, e_A) = I\right\} \cup \left\{e_A : \boldsymbol{\sigma}_A(\theta^H, e_A) = I\right\}$$

has positive Lebesgue measure. Let μ_B denote the probability measure on $T_A^e = [0,1]$ induced by cdf F, an exogenous marginal belief of player B. To ease notation, let

$$\begin{aligned} E_A^{HI} &= \left\{e_A : \boldsymbol{\sigma}_A(\theta^H, e_A) = I\right\}, \\ E_A^{LI} &= \left\{e_A : \boldsymbol{\sigma}_A(\theta^L, e_A) = I\right\}. \end{aligned}$$

A positive fraction of A-types go In and μ_B has full support, therefore $\mu_B[E_A^{LI}] + \mu_B[E_A^{HI}] > 0$. Hence each epistemic type $e_B \in (0,1)$ expects A to go In with positive probability:

$$\begin{aligned} \boldsymbol{\alpha}_B(e_B) &= \boldsymbol{\tau}_B(e_B)(\boldsymbol{\sigma}_A = I | \boldsymbol{\vartheta}_A = \theta^L)\boldsymbol{\tau}_B(e_B)(\boldsymbol{\vartheta}_A = \theta^L) + \boldsymbol{\tau}_B(e_B)(\boldsymbol{\sigma}_A = I | \boldsymbol{\vartheta}_A = \theta^H)\boldsymbol{\tau}_B(e_B)(\boldsymbol{\vartheta}_A = \theta^H) \\ &= \mu_B[E_A^{LI}](1 - e_B) + \mu_B[E_A^{HI}]e_B > 0. \end{aligned}$$

Therefore, for each type $(\theta_B, e_B) \in \Theta \times (0,1)$

$$\mathbb{E}_{(\theta_B, e_B)}[\mathbf{m}_B] = 1 \cdot (1 - \boldsymbol{\alpha}_B(e_B)) + 2 \cdot \boldsymbol{\alpha}_B(e_B) > 1.$$

Since $\mu_A[(0,1)] = 1$, for each e_A

$$\bar{\boldsymbol{\beta}}_A(e_A) = \mathbb{E}_{e_A}\left[\max\{0, \mathbb{E}_{(\boldsymbol{\vartheta}_B, \mathbf{e}_B)}[\mathbf{m}_B] - 1\}\right] > 0.$$

□

Claim 15 *According to B's beliefs, a high-guilt A is strictly more likely to go In than a low-guilt A: $\mu_B[E_A^{HI}] > \mu_B[E_A^{LI}]$. Furthermore, whenever the conditional expectations $\mathbb{E}_{e_B}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right]$ and $\mathbb{E}_{e_B}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right]$ are well defined, they are independent of e_B and satisfy*

$$\mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right] = \frac{1}{\mu_B[E_A^{HI}]} \int_{E_A^{HI}} e_A \mathrm{d}\mu_B(e_A) <$$

$$< \frac{1}{\mu_B[E_A^{LI}]} \int_{E_A^{LI}} e_A \mathrm{d}\mu_B(e_A) = \mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right].$$

Proof $\boldsymbol{\sigma}_A(\theta^L, e_A) = I$ iff $2\mu_A[E_B^{HS}]e_A > 1$, and $\boldsymbol{\sigma}_A(\theta^H, e_A) = I$ iff $2\mu_A[E_B^{HS}]e_A > 1 - \theta^H \bar{\boldsymbol{\beta}}_A(e_A)$. Note that $\theta^H \bar{\boldsymbol{\beta}}_A(e_A) > 0$ because $\theta^H > 0$ by assumption and $\bar{\boldsymbol{\beta}}_A(e_A) > 0$ by Claim 14. Since μ_B has full support,

$$\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}] = \mu_B[\{e_A : 1 - \theta^H \bar{\boldsymbol{\beta}}_A(e_A) < 2\mu_A[E_B^{HS}]e_A \leq 1\}] > 0.$$

Recall that, according to B's beliefs, $\boldsymbol{\vartheta}_A$ and $\mathbf{e}_A$ are independent. Therefore, for every $x \in [0,1]$,

$$\mathbb{P}_{e_B}[\mathbf{e}_A < x | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L] = \mathbb{P}_{e_B}[\mathbf{e}_A < x | \mathbf{e}_A \in E_A^{LI} \cap \boldsymbol{\vartheta}_A = \theta^L] = \mathbb{P}[\mathbf{e}_A < x | \mathbf{e}_A \in E_A^{LI}]$$

whenever the conditional probability is well defined (that is, for $\mu_B[E_A^{LI}] > 0$ and $e_B < 1$). The conditional probability $\mathbb{P}[\mathbf{e}_A < x | \mathbf{e}_A \in E_A^{LI}]$ is independent of e_B because it is determined by the common marginal belief μ_B on $T_A^e = [0,1]$ generated by cdf F:

$$\mathbb{P}[\mathbf{e}_A < x | \mathbf{e}_A \in E_A^{LI}] = \frac{\mu_B[\{e_A \in E_A^{LI} : e_A < x\}]}{\mu_B[E_A^{LI}]}.$$

Similarly,

$$\mathbb{P}_{e_B}[\mathbf{e}_A < x | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H] = \mathbb{P}[\mathbf{e}_A < x | \mathbf{e}_A \in E_A^{HI}] = \frac{\mu_B[\{e_A \in E_A^{HI} : e_A < x\}]}{\mu_B[E_A^{HI}]}$$

whenever the conditional probability is well defined (that is, for $e_B > 0$, since we know that $\mu_B[E_A^{HI}] > 0$). Notice that $E_A^{LI} = (\hat{e}_A^L, 1] \subset E_A^{HI} \subseteq [0,1]$. Therefore, for each $e_B \in (0,1)$,

$$\begin{aligned} & \mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right] \\ = & \frac{1}{\mu_B[E_A^{HI}]} \int_{E_A^{HI}} e_A \mathrm{d}\mu_B(e_A) < \frac{1}{\mu_B[E_A^{LI}]} \int_{E_A^{LI}} e_A \mathrm{d}\mu_B(e_A) \\ = & \mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right] \end{aligned}$$

where the second conditional expectation is well defined if $\mu_B[E_A^{LI}] > 0$, i.e. if $\hat{e}_A^L < 1$. □

Claim 16 *The first-order endogenous belief of B is*

$$\boldsymbol{\alpha}_B(e_B) = \mu_B[E_A^{LI}] + e_B\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right),$$

which is strictly increasing in e_B.

Proof The first-order endogenous belief of B is

$$\begin{aligned} \boldsymbol{\alpha}_B(e_B) &= \mathbb{P}[\boldsymbol{\sigma}_A = I] = \mu_B[E_A^{LI}](1 - e_B) + \mu_B[E_A^{HI}]e_B \\ &= \mu_B[E_A^{LI}] + e_B\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right). \end{aligned}$$

Notice that $\boldsymbol{\alpha}_B$ is strictly increasing in e_B given that $\mu_B[E_A^{HI}] > \mu_B[E_A^{LI}]$, as shown in Claim 15. □

Claim 17 *The second-order endogenous belief of B is such that*

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = \mu_A[E_B^{HS}]\mathbb{E}(\mathbf{e}_A),$$

which is constant, and

$$\boldsymbol{\beta}_B^{I}(e_B) = \mu_A[E_B^{HS}]\left(\mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right](1 - e_B) + \mathbb{E}\left[\mathbf{e}_A | \boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right] e_B\right),$$

which is decreasing (strictly, if $\mu_A[E_B^{HS}] > 0$). Moreover

$$\boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = \begin{cases} S, & \text{if } e_B < \hat{e}_B^H, \\ K, & \text{otherwise,} \end{cases}$$

where $\hat{e}_B^H$ satisfies the incentive conditions

$$\begin{aligned} \hat{e}_B^H = 0 &\implies \boldsymbol{\beta}_B^I(\hat{e}_B^H) \leq \frac{1}{\theta^H}, \\ \hat{e}_B^H \in (0,1) &\implies \boldsymbol{\beta}_B^I(\hat{e}_B^H) = \frac{1}{\theta^H}, \\ \hat{e}_B^H = 1 &\implies \boldsymbol{\beta}_B^I(\hat{e}_B^H) \geq \frac{1}{\theta^H}. \end{aligned}$$

Proof The second-order belief of B on $\boldsymbol{\alpha}_A$ is independent of e_B because, by assumption, $\boldsymbol{\alpha}_A$ depends only on e_A and each type of B has the same marginal belief μ_B (the measure generated by cdf F) on $T_B^e = [0,1]$. Specifically,

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A] = \mathbb{E}_{e_B}(\mu_A[E_B^{HS}]\mathbf{e}_A) = \mu_A[E_B^{HS}]\mathbb{E}(\mathbf{e}_A).$$

Given that $\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A|\boldsymbol{\sigma}_A = I]$ and using Claims 13 and 15, we obtain

$$\begin{aligned}\boldsymbol{\beta}_B^I(e_B) &= \mu_A[E_B^{HS}]\mathbb{E}_{e_B}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I\right] \\ &= \mu_A[E_B^{HS}]\left(\mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right](1-e_B) + \mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right]e_B\right)\end{aligned}$$

if the denominator is positive. Therefore $\boldsymbol{\beta}_B^I(\cdot)$ is decreasing in e_B, given that

$$\frac{\partial \boldsymbol{\beta}_B^I(e_B)}{\partial e_B} = \mu_A[E_B^{HS}]\left(\mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^H\right] - \mathbb{E}\left[\mathbf{e}_A|\boldsymbol{\sigma}_A = I \cap \boldsymbol{\vartheta}_A = \theta^L\right]\right) \leq 0$$

by Claim 15 (note that $\mu_A[E_B^{HS}]$ may be zero). The incentive condition for the high-guilt type of B implies that he chooses S iff $\boldsymbol{\beta}_B^I(e_B) > \frac{1}{\theta^H}$. Therefore B's decision function is weakly decreasing in e_B:

$$\boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = \begin{cases} S, & \text{if } e_B < \hat{e}_B^H, \\ K, & \text{otherwise,} \end{cases}$$

where $\hat{e}_B^H$ satisfies the following conditions

$$\begin{aligned}\hat{e}_B^H &= 0 \Rightarrow \boldsymbol{\beta}_B^I(0) \leq \frac{1}{\theta^H}, \\ \hat{e}_B^H &\in (0,1) \Rightarrow \boldsymbol{\beta}_B^I(\hat{e}_B^H) = \frac{1}{\theta^H}, \\ \hat{e}_B^H &= 1 \Rightarrow \boldsymbol{\beta}_B^I(1) \geq \frac{1}{\theta^H}.\end{aligned}$$

□

Claim 18 *The second-order endogenous belief of A is such that*

$$\begin{aligned}\bar{\boldsymbol{\beta}}_A(e_A) &= \mu_B[E_A^{LI}]\left(3 - 2\mu_A[E_B^{HS}]e_A\right) + 3\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B\right] \\ &\quad -2\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)e_A\mu_A[E_B^{HS}]\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right];\end{aligned}$$

moreover

$$\boldsymbol{\sigma}_A\left(\theta^H, e_A\right) = \begin{cases} I, & \textit{if } e_A \geq \hat{e}_A^H, \\ O, & \textit{otherwise,} \end{cases}$$

where $\hat{e}_A^H \in [0,1)$ satisfies the following incentive conditions

$$\begin{aligned}\hat{e}_A^H &= 0 \Rightarrow 1 - \theta^H\bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) \leq 2\boldsymbol{\alpha}_A\left(e_A\right), \\ \hat{e}_A^H &> 0 \Rightarrow 1 - \theta^H\bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) = 2\boldsymbol{\alpha}_A\left(e_A\right).\end{aligned}$$

Proof Remember that B's disappointment depends on whether B plans to choose S or K after I. Therefore also $\bar{\boldsymbol{\beta}}_A(e_A)$ depends on whether A expects B to choose S or K, as follows:

$$
\begin{aligned}
\bar{\boldsymbol{\beta}}_A(e_A) &= \mathbb{E}_{e_A}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K\right]\mathbb{P}_{e_A}\left[\boldsymbol{\sigma}_B = K\right] + \mathbb{E}_{e_A}\left[\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = S\right]\mathbb{P}_{e_A}\left[\boldsymbol{\sigma}_B = S\right] \\
&= \mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^L\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = K|\boldsymbol{\vartheta}_B = \theta^L\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^L\right] \\
&\quad +\mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = K|\boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^H\right] \\
&\quad +\mathbb{E}\left[\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^H\right] \\
&= \mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\vartheta}_B = \theta^L\right](1-e_A) + \mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right]e_A\left(1-\mu_A[E_B^{HS}]\right) \\
&\quad +\mathbb{E}\left[\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]e_A\mu_A[E_B^{HS}] \\
&= 3\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B\right]\right)(1-e_A) \\
&\quad +3e_A\left(1-\mu_A[E_B^{HS}]\right)\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \\
&\quad +e_A\mu_A[E_B^{HS}]\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \\
&= \mu_B[E_A^{LI}]\left(3\left(1-e_A\right) + 3\left(1-\mu_A[E_B^{HS}]\right)e_A + e_A\mu_A[E_B^{HS}]\right) \\
&\quad +\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)3\mathbb{E}\left[\mathbf{e}_B\right](1-e_A) \\
&\quad +\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)e_A 3\left(1-\mu_A[E_B^{HS}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right] \\
&\quad +\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)e_A\mu_A[E_B^{HS}]\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right] \\
&= \mu_B[E_A^{LI}]\left(3 - 2\mu_A[E_B^{HS}]e_A\right) + 3\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B\right] \\
&\quad -2\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)e_A\mu_A[E_B^{HS}]\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right],
\end{aligned}
$$

where the second equality is a decomposition of the expected value that takes into account that $\mathbb{P}\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^L\right] = 0$; in the third equality we replace probabilities with their specific expressions; the fourth equality is obtained replacing

$$
\boldsymbol{\alpha}_B(e_B) = \mu_B[E_A^{LI}] + e_B\left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right);
$$

finally, the last equality makes use of the following equivalence relation between the expected values (conditional and unconditional) of $\mathbf{e}_B$, based on the independence of $\mathbf{e}_B$ and $\boldsymbol{\vartheta}_B$

$$
\begin{aligned}
\mathbb{E}_{e_A}\left[\mathbf{e}_B\right] &= \mathbb{E}_{e_A}\left[\mathbf{e}_B|\boldsymbol{\vartheta}_B = \theta^H\right] \\
&= \left(1-\mu_A[E_B^{HS}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right] + \mu_A[E_B^{HS}]\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right].
\end{aligned}
$$

We can therefore conclude that $\bar{\boldsymbol{\beta}}_A(\cdot)$ is decreasing in e_A given that

$$
\frac{\partial\bar{\boldsymbol{\beta}}_A}{\partial e_A} = -2\mu_A[E_B^{HS}]\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \leq 0.
$$

From A's incentive condition we can see that type (θ^H, e_A) of A chooses I when

$$
2\mu_A[E_B^{HS}]e_A + \theta^H\bar{\boldsymbol{\beta}}_A(e_A) > 1.
$$

Next we show that either (i) the left hand side (LHS) is increasing in e_A, hence $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is increasing or constant, or (ii) the LHS is larger than 1, hence $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is constant at I. Differentiating the LHS and using the expression for $\frac{\partial\bar{\boldsymbol{\beta}}_A}{\partial e_A}$ we obtain:

$$
2\mu_A[E_B^{HS}] + \theta^H\frac{\partial\bar{\boldsymbol{\beta}}_A}{\partial e_A} = 2\mu_A[E_B^{HS}]\left(1 - \theta^H\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right).
$$

Therefore the LHS is increasing iff

$$
\theta^H\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \leq 1.
$$

Suppose the LHS is strictly decreasing, that is

$$\theta^H \left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right) \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) > 1. \tag{11}$$

Note that by Claim 17 $\mathbb{E}[\mathbf{e}_B] \geq \mathbb{E}[\mathbf{e}_B | \mathbf{e}_B < \hat{e}_B^H] = \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]$; therefore the following inequalities, which imply that the LHS is larger than 1 (and hence $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is constant at I), hold:

$$\theta^H \bar{\boldsymbol{\beta}}_A(e_A) \geq \theta^H \left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right) \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \left(3 - 2\mu_A[E_B^{HS}] e_A\right) > 1.$$

In particular the first inequality holds because the expression for $\bar{\boldsymbol{\beta}}_A(e_A)$ and the fact that $\mathbb{E}[\mathbf{e}_B] \geq \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]$ imply

$$\begin{aligned}
\theta^H \bar{\boldsymbol{\beta}}_A(e_A) &\geq \mu_B[E_A^{LI}] \left(3 - 2\mu_A[E_B^{HS}] e_A\right) + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right) 3\mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right] \\
&\quad -2 \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right) e_A \mu_A^H \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right] \\
&= \left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right) \mathbb{E}\left[\mathbf{e}_B | \boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) \left(3 - 2\mu_A[E_B^{HS}] e_A\right).
\end{aligned}$$

The second inequality holds by eq. (11) and because $\left(3 - 2\mu_A[E_B^{HS}] e_A\right) > 1$.

Therefore

$$\boldsymbol{\sigma}_A\left(\theta^H, e_A\right) = \begin{cases} I, & \text{if } e_A \geq \hat{e}_A^H, \\ O, & \text{otherwise,} \end{cases}$$

where $\hat{e}_A^H \in [0, 1)$ satisfies the incentive condition

$$\begin{aligned}
\hat{e}_A^H &= 0 \Rightarrow 1 - \theta^H \bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) \leq 2\boldsymbol{\alpha}_A\left(e_A\right), \\
\hat{e}_A^H &> 0 \Rightarrow 1 - \theta^H \bar{\boldsymbol{\beta}}_A(\hat{e}_A^H) = 2\boldsymbol{\alpha}_A\left(e_A\right).
\end{aligned}$$

□

Claim 19 *If $\theta^H \leq \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$ all the equilibria are trivial.*

Proof By eq. (10), $\frac{1}{3\mathbb{E}[\mathbf{e}_B]} < 1$. Recall that, by the analog of Claim 5, $\boldsymbol{\sigma}_B\left(\theta^L, e_A\right) = \boldsymbol{\sigma}_B\left(\theta^H, e_A\right) = K$ is B's only optimal strategy when $\theta^H \leq 1$; therefore in this region there are no equilibria in which B chooses S for a set of types with positive measure.

In order to show that all equilibria are trivial if $\theta^H \leq \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$, we proceed by contraposition and show that if $\theta^H \leq 1$ and (despite the fact that all types of B Keep) there is a non-trivial equilibrium, then $\theta^H > \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$.

Suppose that in equilibrium a non-null set of A-types chooses I, given that $\boldsymbol{\sigma}_B\left(\theta_B, e_B\right) = K$ for every $(\theta_B, e_B) \in \{\theta^L, \theta^H\} \times [0, 1]$. In this case A can choose In only if $\boldsymbol{\vartheta}_A = \theta^H$. Type (θ^H, e_A) goes In iff

$$\theta^H > \frac{1}{\bar{\boldsymbol{\beta}}_A\left(e_A\right)},$$

where

$$\bar{\boldsymbol{\beta}}_A\left(e_A\right) = \mathbb{E}_{e_A}\left[3\boldsymbol{\alpha}_B\right] = \mathbb{E}_{e_A}\left[3\mu_B[E_A^{HI}]\mathbf{e}_B\right] = 3\mu_B[E_A^{HI}]\mathbb{E}[\mathbf{e}_B] \leq 3\mathbb{E}[\mathbf{e}_B].$$

Hence there is a non-null set of A-types going In only if

$$\theta^H > \frac{1}{\bar{\boldsymbol{\beta}}_A\left(e_A\right)} \geq \frac{1}{3\mathbb{E}[\mathbf{e}_B]}.$$

for some e_A. □

Claim 20 *If* $\theta^H > \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$ *the following is an equilibrium*

$$\begin{aligned}\boldsymbol{\sigma}_A\left(\theta^L, e_A\right) &= O,\\ \boldsymbol{\sigma}_A\left(\theta^H, e_A\right) &= I,\\ \boldsymbol{\sigma}_B\left(\theta^L, e_B\right) &= \boldsymbol{\sigma}_B\left(\theta^H, e_B\right) = K,\end{aligned}$$

for every e_A, e_B. *This equilibrium induces the endogenous beliefs* $\boldsymbol{\alpha}_A(e_A) = 0$, $\bar{\boldsymbol{\beta}}_A(e_A) = 3\mathbb{E}[\mathbf{e}_B]$, $\boldsymbol{\alpha}_B(e_B) = e_B$, $\boldsymbol{\beta}_B^{\varnothing}(e_B) = \boldsymbol{\beta}_B^{I}(e_B) = 0$. *Moreover, this is the unique non-trivial equilibrium if* $\frac{1}{3\mathbb{E}[\mathbf{e}_B]} < \theta^H \leq 1$.

Proof If $\boldsymbol{\sigma}_B(\theta_B, e_B) = K$ for each type (θ_B, e_B) of B, the first order belief of A is $\boldsymbol{\alpha}_A(e_A) = 0$ and the best response of each type (θ^L, e_A) is to stay Out: $\boldsymbol{\sigma}_A(\theta^L, e_A) = O$ for every e_A. Now consider high-guilt types (θ^H, e_A). The calculations in the proof of Claim 19 show that $\bar{\boldsymbol{\beta}}_A(e_A) = 3\mu_B[E_A^{HI}]\mathbb{E}[\mathbf{e}_B]$. In the candidate equilibrium each high-guilt type of A goes In, therefore $\mu_B[E_A^{HI}] = 1$ and $\bar{\boldsymbol{\beta}}_A(e_A) = 3\mathbb{E}[\mathbf{e}_B]$ for every e_A. Since $\theta^H > \frac{1}{3\mathbb{E}[\mathbf{e}_B]}$, the incentive condition for $\boldsymbol{\sigma}_A(\theta^H, e_A) = I$ is always satisfied:

$$\bar{\boldsymbol{\beta}}_A(e_A)\,\theta^H = 3\mathbb{E}[\mathbf{e}_B]\theta^H > 1.$$

Since $\boldsymbol{\alpha}_A(e_A) = 0$, then $\boldsymbol{\beta}_B^{\varnothing}(e_B) = \boldsymbol{\beta}_B^{I}(e_B) = 0$ and the best response of each type (θ_B, e_B) is indeed to Keep.

Finally, by the analog of Claim 5, $\boldsymbol{\sigma}_B = K$ is B's only equilibrium decision function if $\theta^H \leq 1$, therefore the equilibrium described above is the unique equilibrium for $\theta^H \in \left(\frac{1}{3\mathbb{E}[e_B]}, 1\right]$. □

■

References

[1] Attanasi G., P. Battigalli and R. Nagel (2012). "Disclosure of Belief-Dependent Preferences in the Trust Game," typescript.

[2] Attanasi G. and R. Nagel (2008). "A survey of psychological games: theoretical findings and experimental evidence," in: A. Innocenti and P. Sbriglia (Eds.), *Games, Rationality and Behavior. Essays on Behavioral Game Theory and Experiments.* Houndmills: Palgrave McMillan, 204-232.

[3] Battigalli P. and M. Dufwenberg (2007). "Guilt in Games," *American Economic Review, Papers and Proceedings*, 97, 170-176.

[4] Battigalli P. and M. Dufwenberg (2009). "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1-35.

[5] Battigalli P. and M. Siniscalchi (2003). "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, 3 (1), Article 3.

[6] Battigalli P., G. Charness and M. Dufwenberg (2012). "Deception: The Role of Guilt," IGIER working paper 457 (forthcoming *Journal of Economic Behavior and Organization).*

[7] Bellemare C., A. Sebald and M. Strobel (2011). "Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models," *Journal of Applied Econometrics*, 26, 437-453.

[8] Berg J., J. Dickhaut and K. McCabe (1995). "Trust, Reciprocity, and Social-History," *Games and Economic Behavior*, 10, 122-142.

[9] Binmore, K., J. Gale and L. Samuelson (1995). "Learning To Be Imperfect: The Ultimatum Game," *Games and Economic Behavior,* 8, 56-90.

[10] Brandenburger A. and E. Dekel (1987). "Rationalizability and Correlated Equilibria," *Econometrica*, 55**,** 1391-1402.

[11] Buskens V. and W. Raub (2008). "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust," in R. Wittek, T.A.B. Snijders & V. Nee (Eds.) *Handbook of Rational Choice Social Research*, New York: Russell Sage.

[12] Caplin A. and J. Leahy (2004). "The supply of information by a concerned expert," *Economic Journal*, 114, 487-505.

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

[14] Charness G. and M. Dufwenberg (2006). "Promises and Partnership," *Econometrica,* 74, 1579-1601.

[15] Cooper D. and J. Kagel (2013). "Other-Regarding Preferences: A Selective Survey of Experimental Results," to appear in J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics,* 2. Princeton: Princeton University Press.

[16] Dekel E., D. Fudenberg and D. Levine (2004). "Learning to Play Bayesian Games," *Games and Economic Behavior*, 46, 282-303.

[17] Dufwenberg M. (2006). "Psychological Games," in S.N. Durlauf and L.E. Blume (Eds.), *The New Palgrave Dictionary of Economics,* 6, 714-718.

[18] Dufwenberg M. (2002). "Marital investment, time consistency and emotions," *Journal of Economic Behavior and Organization*, 48, 57-69.

[19] Dufwenberg M. and U. Gneezy (2000). "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior*, 30, 163-182.

[20] Ellingsen T., M. Johannesson, S. Tjotta and G. Torsvik (2010), "Testing Guilt Aversion," *Games and Economic Behavior*, 68, 95-107.

[21] Esponda I. (2012). "Rationalizable Conjectural Equilibrium: A Framework for Robust Predictions," *Theoretical Economics,* forthcoming.

[22] Gachter S., D. Nosenzo, E. Renner and M. Sefton, (2012). "Who makes a good leader? Cooperativeness, optimism and leading-by-example," *Economic Inquiry*, 50, 867-879.

[23] Geanakoplos J., D. Pearce and E. Stacchetti (1989). "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60-79.

[24] Gneezy U. (2005), "Deception: The Role of Consequences," *American Economic Review,* 95, 384-394.

[25] Guerra G. and D.J. Zizzo (2004). "Trust responsiveness and beliefs," *Journal of Economic Behavior and Organization*, 55, 25-30.

[26] Harsanyi J. (1967-68). "Games of incomplete information played by Bayesian players. Parts I, II, III," *Management Science*, 14, 159-182, 320-334, 486-502.

[27] Haselton M.G. and T. Ketelaar (2006). "Irrational Emotions or Emotional Wisdom? The Evolutionary Psychology of Emotions and Behavior," in J. Forgas (Ed.) *Hearts and Minds: Affective Influences on Social Cognition and Behavior.* (Frontiers of Psychology Series). New York: Psychology Press.

[28] Ross L., D. Greene, and P. House (1977). "The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Processes," *Journal of Experimental Social Psychology*, 13, 279-301.

[29] Reuben E., P. Sapienza and L. Zingales (2009). "Is Mistrust Self-Fulfilling?" *Economic Letters*, 104, 89-91.

[30] Shaked M, and J.G. Shantikumar (2007). *Stochastic Orders.* New York: Springer.

[31] Tadelis S. (2011): "The Power of Shame and the Rationality of Trust," typescript, UC Berkeley.

[32] Vanberg C. (2008), "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica,* 76, 1467-1480.

[33] Weibull J.W. (1996): "The mass action interpretation," in H. Kuhn et al "The work of John Nash in game theory", *Journal of Economic Theory*, 69, 153-185.