INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Incomplete Information Models of Guilt Aversion in the Trust Game

Giuseppe Attanasi, Pierpaolo Battigalli and Elena Manzoni

Working Paper n. 480

This Version: January, 2015

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.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 guilt aversion in the Trust Game. In our models, agents have heterogeneous belief hierarchies. We characterize equilibria where trust occurs with positive probability. 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 (investment) 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 2013, 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. Charness & Dufwenberg (2006) provide 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

*We thank two anonymous referees and the anonymous Associate Editor for helpful comments conducive to a substantially improved draft. We also thank Nicodemo De Vito, Martin Dufwenberg, Alejandro Francetich, Astrid Gamba, Arsen Palestini, Cintia Retz Lucci, Marco Scarsini, Severine Toussaert, and seminar participants in Toulouse, the Catholic University of Milan, Politecnico of Milan, and Università La Sapienza in Rome for their comments. Giuseppe Attanasi gratefully acknowledges financial support of ERC starting grant DU 283953 and of "Attractivité" IDEX 2013 (University of Strasbourg). Pierpaolo Battigalli gratefully acknowledges financial support of ERC advanced grant 324219. Elena Manzoni gratefully acknowledges financial support of PRIN 2010-2011 "New Approaches to political economy: positive political theories, empirical evidence and experiments in laboratory."

work corroborates this hypothesis (e.g., Reuben *et al.* 2009, Bellemare *et al.* 2011, Chang *et al.* 2011, Attanasi *et al.* 2013).[1]

Of course, when subjects' preferences 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 material payoffs at terminal nodes) are made common knowledge in the experiment. 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 the Bayesian equilibria (Harsanyi, 1967-68) of 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] A key feature of both models is that agents playing in a given role hold heterogeneous beliefs about the type of the co-player, which implies heterogeneous first- and second-order beliefs about actions. In the first model it is common knowledge that player A, the "truster", is selfish and only player B, the "trustee", can feel guilt. In the second model, instead, guilt sensitivity and beliefs about it do not depend on the role played in the game (player A or player B). The first model is more tractable and it may be appropriate for situations where the players come from different populations, e.g. when A is a firm and B is a worker. 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 resonates with (i) the evolutionary psychology of emotions (e.g., Haselton & Ketelaar 2006), which suggests that, when a single emotion (guilt) operates in a variety of different domains, its effects are moderated by contextual cues; and with (ii) the conceptual act theory of emotion (e.g., Barrett 2006), which posits that people experience an (in our case, anticipated) emotion by categorizing an instance of affective feeling (anticipated disappointment of the other); with this, it is plausible that the role played in the interaction is part of the categorization process. Finally, the different responses to oxytocin of trusters and trustees (e.g., Kosfeld *et al.* 2005, Zak *et al.* 2005), and findings in the animal literature on the reactivity of oxytocin to social cues (e.g., Carter & Keverne 2002) provide some indirect evidence supporting the role-dependent model of guilt aversion.[3]

Our approach 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 the first paper offering a fully-fledged Bayesian equilibrium analysis of

[1] See also Dufwenberg & Gneezy (2000) and Guerra & Zizzo (2004). Charness & Dufwenberg (2011) find support for the "guilt-from-blame" model of Battigalli & Dufwenberg (2007), which assumes that i anticipates guilt if he thinks that j is going to blame i for letting him down. Vanberg (2008) and Ellingsen *et al.* (2010) question the guilt-aversion interpretation of pro-social behavior in the Trust Game.

[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] On the one hand, higher levels of oxytocin in trustees are correlated (between subjects) with higher investment by trusters and higher sharing (Zak *et al.* 2005, see also Zak 2008). According to the guilt-aversion model, oxytocin can thus be interpreted as the transmitter from second-order beliefs (expected disappointment from not giving) to the pro-social action. But in the case of trusters, higher levels of oxytocin are *not* correlated (between subjects) with higher investment. And yet, Kosfeld *et al.* (2005) show that an exogenously induced increase in the oxytocin level increases investment, which suggests that it is the difference between baseline and actual level of oxytocin that affects pro-social behavior. We submit that the role of truster is a social cue that shuts down the link between second-order beliefs and oxytocin levels (cf. Carter & Keverne 2002); thus, oxytocin differences between trusters mainly reflect differences in baseline levels, not in second-order beliefs.

guilt aversion.[4]

Our paper is related to Attanasi *et al.* (2013), 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 a 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 organize the data of their experiment.

Also, under mild assumptions about the empirical distribution of types, our analysis implies the positive correlation between the second-order beliefs and pro-social action of B-subjects found by Charness & Dufwenberg (2006). In the final discussion (Section 6), we comment extensively on such correlation, and – more generally – on the relevance of our models for experiments.

As argued above, the assumption that interacting individuals have belief-dependent social preferences naturally leads to an incomplete-information analysis. Therefore, 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. Indeed, we present the more abstract methodological material on Bayesian equilibrium and psychological games in Section 3 so that it can be easily extended and applied to different games.

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 a model with role-dependent guilt, where A is known to be selfish. Section 5 puts forward and analyzes a model with role-independent guilt. Section 6 concludes with a discussion of the empirical implications of our analysis and of the situations where it can be usefully applied. 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 material payoffs depicted in Figure 1.

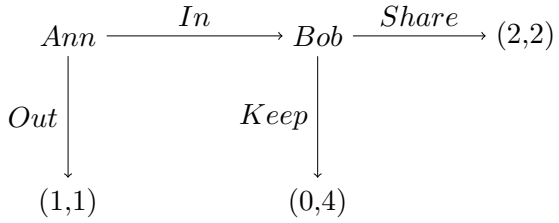

Figure 1: The Trust Minigame form with material payoffs

[4]Some papers analyze incomplete-information models of games with belief-dependent preferences; see, for example, Caplin & Leahy (2004), Ong (2011) and Tadelis (2011). Unlike ours, none of these models features heterogenous beliefs, which are instead allowed for by Battigalli *et al.* (2013). The latter paper analyzes the cheap-talk game of Gneezy's (2005) experiment under the assumption that the sender is affected by an unknown sensitivity to guilt. This analysis is based on two rounds of elimination of non-best replies under mild assumptions about heterogeneous beliefs.

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 of guilt feelings on 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. 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 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 material payoff distributions represented by the following psychological utility function:

$$u_i = m_i - \theta_i \max\{0, \mathbb{E}_{-i}[\mathbf{m}_{-i}] - m_{-i}\}, \tag{1}$$

where m_i is the material payoff of i, $\theta_i \geq 0$ is his guilt sensitivity, and $\max\{0, \mathbb{E}_{-i}[\mathbf{m}_{-i}] - m_{-i}\}$ measures the extent of the co-player's disappointment given the co-player's 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 by the following parametrized psychological game:

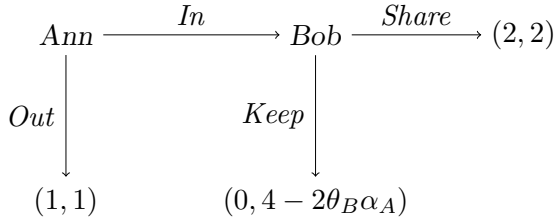

Figure 2: The Trust Minigame with psychological utilities

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

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

[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 material payoff of A.

where $m_i(z)$ denotes the material payoff of i at terminal history $z \in \{O, (I,K), (I,S)\}$. Thus, the psychological utility at $z = (I,K)$ of 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:

$$\mathbb{E}_B^S[\mathbf{u}_B] - \mathbb{E}_B^K[\mathbf{u}_B] = \mathbb{E}_B[u_B(I, S, \boldsymbol{\alpha}_A) - u_B(I, K, \boldsymbol{\alpha}_A)|I] \cdot \mathbb{P}_B[I] = 2(\theta_B\beta_B^I - 1)\alpha_B. \qquad (2)$$

By definition, $0 \leq \beta_B^I \leq 1$; thus, in an equilibrium with $\alpha_B > 0$, Bob Shares if $\alpha_B > 0$, $\theta_B > 1$ and $\beta_B^I > \frac{1}{\theta_B}$.

The assumption that only player B may be sensitive to guilt is removed in Section 5, where we analyze a model with role-independent guilt.

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 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. The main concepts are illustrated by a leading example.

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.

3.1 Type structures

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. Formally, the **psychological utility** of i is a parametrized function

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

where Z is the set of terminal histories (play paths) of the game, and $\mathcal{H}_i$ ($\mathcal{H}_{-i}$) is a space of endogenous hierarchical beliefs of player i ($-i$).[6] Since in our applications θ_i is the guilt sensitivity parameter of player i, we call θ_i **"guilt type."** When the parameter set Θ_i is a singleton, the

[6] See Geanakoplos *et al.* (1989) and Battigalli & Dufwenberg (2009). In the latter, $\mathcal{H}_i$ is a space of hierarchical *conditional* beliefs.

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, i.e. $\Theta_A = \{0\}$.

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

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

Elements of T_i are called Harsanyi types, or simply **types**. A Harsanyi type specifies both the guilt type (more generally, the utility function) and the exogenous beliefs of player i.[7] Note that we use bold symbols to denote functions interpreted as **random variables**, that is, functions that depend on the **state of the world** (t_A, t_B). Function $\boldsymbol{\vartheta}_i$ specifies the psychological utility (guilt sensitivity) of type t_i, and function $\boldsymbol{\tau}_i$ determines the exogenous beliefs of t_i about the utility and beliefs of the co-player $-i$. In particular, we explain below how each type t_i in the structure determines a whole hierarchy of exogenous beliefs for player i. 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 $\boldsymbol{\vartheta}_i^{-1}([0, x])$ and $\boldsymbol{\vartheta}_i \leq x$ denote the set $\{t_i : \boldsymbol{\vartheta}_i(t_i) \leq x\}$, that is, the event that the guilt type of i is at most x. We use whatever notation is more convenient and transparent in the given context.

It may be assumed without essential loss of generality that the set of types is a Cartesian product $T_i = \Theta_i \times \mathcal{E}_i$, so that a type is a pair (θ_i, e_i), and that beliefs about the co-player's type are determined only by the second element e_i, also called **epistemic type**. In this case, function $\boldsymbol{\vartheta}_i$ is the projection map $(\theta_i, e_i) \mapsto \boldsymbol{\vartheta}_i(\theta_i, e_i) = \theta_i$, and function $\boldsymbol{\tau}_i$ depends only on e_i, hence it makes sense to write $\boldsymbol{\tau}_i(e_i)$.

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

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

In this paper, we focus on Bayesian psychological games based on the Trust Minigame with guilt aversion, that is, the game form of Figure 1 with parametrized utility functions given by eq. (1).

3.2 Higher-order exogenous beliefs

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

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

For example, $\mathbf{p}_A^1(t_A)\,[\{t_B : \boldsymbol{\vartheta}_B(t_B) > 2\}]$ is the subjective probability assigned by type t_A to the event that the guilt sensitivity of B is more than 2. We can write this probability more compactly as $\mathbf{p}_A^1(t_A)\,[\boldsymbol{\vartheta}_B > 2]$.

With this, 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

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

For example, $\mathbf{p}_B^2(t_B)\left[\left\{t_A : p_A^1(t_A)\,[\boldsymbol{\vartheta}_B > 2] \geq \frac{1}{2}\right\}\right]$ is the subjective probability assigned by type t_B to the event that A believes that $\boldsymbol{\vartheta}_B > 2$ is at least as likely as $\boldsymbol{\vartheta}_B \leq 2$.

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.

[7] All our models satisfy the following 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 $\boldsymbol{\vartheta}_i$, $\boldsymbol{\tau}_i$ are *continuous*. This implies that the sets and functions we consider satisfy the necessary measurability requirements.

3.3 Leading example: exogenous beliefs

We illustrate these abstract concepts with an example, which is (essentially) a special case of our model with role-dependent guilt. Suppose that it is common knowledge that A is selfish, whereas B can either have a low guilt type $\theta^L < 1$, or a high guilt type $\theta^H > 2$; therefore, $\Theta_A = \{0\}$ and $\Theta_B = \{\theta^L, \theta^H\}$. The exogenous beliefs of each player i are determined by the epistemic type $e_i \in \mathcal{E}_i$. Since player A has only one possible guilt type, the epistemic and Harsanyi types of A coincide, and we can ease notation writing $e_A = t_A$. There is a continuum of epistemic types on both sides. Specifically, we let $e_A = t_A \in T_A$ parametrize the subjective probability assigned by A to the high-guilt type of B: $t_A = \mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B = \theta^H]$.[8] Therefore, we let $T_A = \mathcal{E}_A = [0,1]$. Furthermore, we assume that the set of possible epistemic types of B is $[0,1]$ as well. This is just a convenient parametrization. Thus $T_B = \Theta_B \times \mathcal{E}_B = \{\theta^L, \theta^H\} \times [0,1]$.While the set of epistemic types of A is easily seen to be isomorphic to the set of exogenous first-order beliefs of A, the meaning of the epistemic types of B can only be understood by considering the belief maps and unraveling the higher-order beliefs corresponding to each type.

The belief maps have the following features. All types of player A believe that the guilt and epistemic type of B are statistically independent; furthermore, they hold the same marginal belief about the epistemic type of B given by a strictly positive density function $f : [0,1] \to \mathbb{R}$, e.g., the uniform distribution $f(e_B) = 1$. On the other hand, different epistemic types of B may hold different beliefs about the type of A: The belief of each e_B is given by a strictly positive density function $f_{e_B} : [0,1] \to \mathbb{R}$. For illustrative purposes, we provide a simple specification of the belief map $e_B \mapsto f_{e_B}$:

$$f_{e_B}(t_A) = \begin{cases} 1 - e_B, & \text{if } 0 \leq t_A \leq \frac{3}{4}, \\ (1 - e_B) + 4e_B, & \text{if } \frac{3}{4} < t_A \leq 1, \end{cases} \tag{3}$$

for all $t_A, e_B \in [0,1]$. In words, f_{e_B} is a mixture of two distributions: epistemic type e_B believes that with probability e_B the type of A is uniformly distributed on $(\frac{3}{4}, 1]$, and with probability $(1 - e_B)$ the type of A is uniformly distributed on $[0,1]$. This implies that higher epistemic types of B hold "higher beliefs" – in the sense of stochastic dominance – about $\mathbf{t}_A$.

To sum up, the belief maps $\boldsymbol{\tau}_i : T_i \to \Delta(T_{-i})$ $(i \in \{A, B\})$ satisfy

$$\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H \cap \mathbf{e}_B \leq y] = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H] \cdot \boldsymbol{\tau}_A(t_A)[\mathbf{e}_B \leq y] = t_A \int_0^y f(e_B) \mathrm{d}e_B,$$

and

$$\boldsymbol{\tau}_B(\theta_B, e_B)[\mathbf{t}_A \leq x] = \int_0^x f_{e_B}(t_A) \mathrm{d}t_A,$$

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

All of the above is assumed to be common knowledge. This gives the type structure $\mathcal{T}$ and a Bayesian psychological game Γ based on the Trust Minigame with guilt aversion.

Exogenous hierarchies of beliefs are relatively simple. In particular, beliefs about the guilt type of A and beliefs about such beliefs are trivial, because A is commonly known to be selfish, whereas the first- and second-order beliefs of, respectively, A and B about the guilt type of B are

$$\mathbf{p}_A^{1,H}(t_A) = \mathbf{p}_A^1(t_A)[\boldsymbol{\vartheta}_B = \theta^H] = t_A$$

and

$$\mathbf{p}_B^2(\theta_B, e_B)\left[\mathbf{p}_A^{1,H} \leq x\right] = \boldsymbol{\tau}_B(\theta_B, e_B)[\mathbf{t}_A \leq x] = \int_0^x f_{e_B}(t_A) \mathrm{d}t_A$$

[8] We often write $\mathbb{P}_{t_A}[\cdot]$ instead of $\boldsymbol{\tau}_A(t_A)[\cdot]$ to ease notation in the context of examples and models.

for all $t_A \in T_A$, $(\theta_B, e_B) \in T_B$, and $x \in [0,1]$. For example, we can use eq. (3) to derive the probability assigned by epistemic type e_B to the event that A deems θ^H at least as likely as θ^L:

$$\mathbf{p}_B^2(\theta_B, e_B)\left[\mathbf{p}_A^{1,H} \geq \frac{1}{2}\right] = \boldsymbol{\tau}_B(\theta_B, e_B)\left[\mathbf{t}_A \geq \frac{1}{2}\right] = \frac{1+e_B}{2}.$$

3.4 Equilibrium

A **Bayesian equilibrium** of the 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 utility, given the endogenous beliefs of type t_i about the co-player's choice and beliefs.[9] Note that, in general, this is a *subjective* notion of equilibrium, because players' exogenous beliefs are not necessarily derived from an objective distribution of types. For more on this, see subsection 3.8.

In a **perfect** Bayesian equilibrium, player B maximizes his *conditional* expected utility upon observing I, with conditional beliefs computed by Bayes' rule, if possible. Of course, if $\boldsymbol{\sigma}_A(t_A) = O$ for every t_A, the conditional second-order belief $\boldsymbol{\beta}_B^I(t_B) = \mathbb{E}_{t_B}[\boldsymbol{\alpha}_A | \boldsymbol{\sigma}_A = I]$ cannot be determined by Bayes' formula, and we cannot rule out the possibility that $\boldsymbol{\beta}_B^I(t_B) = 0$, hence $\boldsymbol{\sigma}_B(t_B) = K$, for every t_B. This in turn implies that each type of A is certain of K $(\boldsymbol{\alpha}_A(t_A) = 0)$, which justifies $\boldsymbol{\sigma}_A(t_A) = O$ for every t_A. This explains the following remark:[10]

Remark 1 *Every Bayesian psychological game based on the Trust Minigame with guilt aversion has a perfect Bayesian equilibrium with $\boldsymbol{\sigma}_A(t_A) = O$ and $\boldsymbol{\sigma}_B(t_B) = K$ for all types t_A and t_B.*

Having established this once and for all, the rest of the analysis is focused on **non-degenerate equilibria** where a positive fraction of A's types choose I. Under our assumptions on exogenous beliefs, this implies that every type of B assigns positive probability to I. As we noticed in Section 2, in this case *ex ante* maximization of psychological utility is equivalent to conditional maximization (see eq. (2)); therefore, *non-degenerate Bayesian equilibria are also perfect.* When A is commonly known to be selfish, such equilibria have another interesting feature: Since the positive fraction of A's types who choose I in equilibrium do this to maximize expected payoff, upon observing I, player B must conclude that A chose I rationally, hence that A assigned at least 50% probability to S. In other words, in a non-degenerate equilibrium, each type t_B is certain conditional on I that $\boldsymbol{\alpha}_A \geq 1/2$: $\boldsymbol{\beta}_B^I(t_B)[\boldsymbol{\alpha}_A \geq 1/2] = 1$. This is exactly the same inference imposed by forward-induction reasoning (see Dufwenberg 2002, and Section 5 of Battigalli & Dufwenberg 2009). Hence, our focus on non-degenerate equilibria can also be motivated as a forward-induction refinement. Indeed, some of our insights are solely based on a kind of step-by-step forward-induction reasoning, while others need the fully-fledged Bayesian equilibrium analysis.

[9] 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."

[10] Remark 1 is an instance of a more general observation: 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 (possibly null) guilt-aversion term, and possibly allowing for incomplete information about guilt parameters. Then, every pure-strategy sequential equilibrium of the material-payoff game G is also a perfect Bayesian equilibrium of the psychological game $\Gamma(G)$. The intuition is that off the equilibrium path players may believe that deviations occurred by mistake and hence do not signal expectations of high material payoffs. Thus, the best reply at off-path information sets is to maximize one's own expected material payoff. See Battigalli & Dufwenberg (2007, Observation 2).

3.5 Higher-order endogenous beliefs

It is important to understand how the type structure and decision functions $\boldsymbol{\sigma}_i$ generate the players' endogenous beliefs. We analyze psychological games where the utility of i (determined by his guilt type θ_i) depends on what $-i$ plans to do ($-i$'s strategy) and on the endogenous first-order beliefs of $-i$. For example, the utility of each 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:

$$\boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B = S]. \tag{4}$$

For player B (and the analyst), $\boldsymbol{\alpha}_A : T_A \to [0,1]$ is a random variable. Type t_B of player B can compute his initial expectation of $\boldsymbol{\alpha}_A$ as follows:[11]

$$\boldsymbol{\beta}_B^{\varnothing}(t_B) = \mathbb{E}_{t_B}[\boldsymbol{\alpha}_A] = \int \boldsymbol{\alpha}_A(t_A)\boldsymbol{\tau}_B(t_B)[\mathrm{d}t_A]. \tag{5}$$

This initial second-order belief reflects how B reasons about the game before playing it.[12] But B takes an action only if he observes I, therefore his choice depends on his second-order belief conditional on I:

$$\boldsymbol{\beta}_B^I(t_B) = \mathbb{E}_{t_B}[\boldsymbol{\alpha}_A|\boldsymbol{\sigma}_A = I] = \frac{1}{\boldsymbol{\alpha}_B(t_B)} \int_{t_A:\boldsymbol{\sigma}_A(t_A)=I} \boldsymbol{\alpha}_A(t_A)\boldsymbol{\tau}_B(t_B)[\mathrm{d}t_A], \text{ if } \boldsymbol{\alpha}_B(t_B) > 0, \tag{6}$$

where $\boldsymbol{\alpha}_B(t_B) = \boldsymbol{\tau}_B(t_B)[\boldsymbol{\sigma}_A = I]$ is the initial endogenous first-order belief of t_B (cf. Section 2).

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 $\mathcal{T}$. But, for the sake of clarity, we will make the key features of endogenous beliefs explicit.

Besides endogenous first- and second-order 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 "material payoff of player i" is[13]

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

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 $\boldsymbol{\sigma}_{-i}(t_{-i})$.

As in the leading example, we use models where the type set of player i can be factorized as $T_i = \Theta_i \times \mathcal{E}_i$, with the convenient parametrization $\mathcal{E}_i = [0,1]$. The second component of $t_i = (\theta_i, e_i)$ – 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 $T_i = \Theta_i \times \mathcal{E}_i$ onto $\mathcal{E}_i = [0,1]$: $\mathbf{e}_i(t_A, t_B) = e_i$ 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 threshold x. Given this, we can ease notation writing the belief maps as function of e_i only, as in $\boldsymbol{\tau}_i(e_i)$, $\boldsymbol{\alpha}_i(e_i)$, $\boldsymbol{\beta}_i(e_i)$.

[11] 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 $\mathbb{E}_{\boldsymbol{\tau}_i(t_i)}[\mathbf{x}_{-i}]$.

[12] Hence it is an interesting feature of beliefs that is worth eliciting in experiments. See the discussion in Section 6.

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

3.6 Leading example: equilibrium and endogenous beliefs

Under the simplifying assumptions of the example (see Section 3.3), there is a unique non-degenerate equilibrium pair of decision functions $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ that can be determined with the forward-induction argument mentioned in Section 3.4: Since in a non-degenerate equilibrium a positive fraction of A-types choose I and each f_{e_B} has full support on $[0,1]$, each epistemic type e_B assigns positive probability to I, that is,

$$\boldsymbol{\alpha}_B(e_B) = \int_{t_A:\boldsymbol{\sigma}_A(t_A)=I} f_{e_B}(t_A)\mathrm{d}t_A > 0.$$

Hence, $\boldsymbol{\beta}_B^I(e_B)$ is determined by Bayes' rule (cf. eq. (6)):

$$\boldsymbol{\beta}_B^I(e_B) = \frac{1}{\boldsymbol{\alpha}_B(e_B)} \int_{t_A:\boldsymbol{\sigma}_A(t_A)=I} \boldsymbol{\alpha}_A(t_A) f_{e_B}(t_A)\mathrm{d}t_A.$$

A's rationality implies[14]

$$\boldsymbol{\sigma}_A(t_A) = \begin{cases} O, & \text{if } \boldsymbol{\alpha}_A(t_A) < \frac{1}{2}, \\ I, & \text{if } \boldsymbol{\alpha}_A(t_A) > \frac{1}{2}. \end{cases}$$

Therefore,

$$\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}\left[\boldsymbol{\alpha}_A | \boldsymbol{\alpha}_A \geq \frac{1}{2}\right] \geq \frac{1}{2}.$$

B's rationality implies

$$\boldsymbol{\sigma}_B(\theta_B, e_B) = \begin{cases} K, & \text{if } \theta_B \boldsymbol{\beta}_B^I(e_B) < 1, \\ S, & \text{if } \theta_B \boldsymbol{\beta}_B^I(e_B) > 1. \end{cases}$$

Since $\theta^L < 1$, $\theta^H > 2$ and $\boldsymbol{\beta}_B^I(e_B) \geq 1/2$, the choice of B is independent of the epistemic type e_B: B Keeps if selfish and Shares if prone to guilt feelings, that is,

$$\boldsymbol{\sigma}_B(\theta_B, e_B) = \begin{cases} K, & \text{if } \theta_B = \theta^L, \\ S, & \text{if } \theta_B = \theta^H. \end{cases}$$

Therefore, the endogenous first-order belief of a type t_A coincides with the exogenous one, that is, the probability assigned to the high-guilt type of B:

$$\boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B = \theta^H] = t_A,$$

and the decision function of A is

$$\boldsymbol{\sigma}_A(t_A) = \begin{cases} O, & \text{if } t_A < \frac{1}{2}, \\ I, & \text{if } t_A > \frac{1}{2}. \end{cases}$$

Although in this example the fine details of higher-order beliefs do not matter, we derive the endogenous second-order beliefs of B for illustrative purposes:

$$\boldsymbol{\beta}_B^{\varnothing}(e_B) = \mathbb{E}_{e_B}[\mathbf{t}_A] = \int_0^1 t_A f_{e_B}(t_A)\mathrm{d}t_A,$$

$$\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}\left[\mathbf{t}_A | \mathbf{t}_A > \frac{1}{2}\right] = \frac{\int_{\frac{1}{2}}^1 t_A f_{e_B}(t_A)\mathrm{d}t_A}{\int_{\frac{1}{2}}^1 f_{e_B}(t_A)\mathrm{d}t_A} = \frac{3+4e_B}{4(1+e_B)},$$

where the last equality follows from eq. (3). Notice that the conditional second-order belief $\boldsymbol{\beta}_B^I(e_B)$ is increasing in e_B. Such monotonicity is exploited in the equilibrium analysis of the general model with many guilt types of Section 4.

[14] We can ignore knife-edge cases because beliefs about the co-player's type are absolutely continuous.

3.7 Harsanyi's method and psychological games

There is a noteworthy difference between the definition of equilibrium in psychological games with complete information and in Bayesian psychological games: In the former, it is necessary to assume that endogenous beliefs of all orders are correct (see Geanakoplos *et al.* 1989, and Battigalli & Dufwenberg 2009); in the latter, it is instead assumed that conjectures about the co-players' decision functions are correct, but there is no explicit condition concerning belief hierarchies. In other words, the analysis of Bayesian psychological games just requires to apply the equilibrium concept that was already on the shelves of standard game theory, whereas an extension of the traditional definition of equilibrium is needed for the analysis of complete-information psychological games. How can the two definitions be reconciled?

Following the method proposed by Harsanyi in his three-part article in *Management Science* (1967-68), we posit a Bayesian game Γ and a profile of decision functions $\boldsymbol{\sigma}$. The pair $(\Gamma, \boldsymbol{\sigma})$ is an interactive beliefs structure generating a description of the possible hierarchies of beliefs about parameters (utility functions) and about choices, as illustrated in the leading example. These hierarchies satisfy by construction all the necessary coherence conditions, given $\boldsymbol{\sigma}$. In a psychological game, if the profile of decision functions $\boldsymbol{\sigma}$ is a Bayesian equilibrium, the choice of each type is a best reply to his belief *hierarchy*, and this is common belief; therefore, no further condition has to be added to the definition of equilibrium. Unlike a Bayesian psychological game, a complete-information psychological game does not come equipped with a type structure, hence there seems to be no way to automatically unfold belief hierarchies starting from a profile of strategies s.

Despite this, there is no substantial difference between the two equilibrium concepts; as a matter of fact, one is a special case of the other. To see this, note that a complete-information psychological game G can be interpreted as a trivial Bayesian game Γ with just one type for each player. In this case, a profile $\boldsymbol{\sigma}$ of maps from types to strategies is just a profile of strategies s in G. Hence, the unique type of each player i corresponds to the degenerate hierarchy of beliefs whereby beliefs of all orders are correct: the first-order belief of i assigns probability one to s_{-i}, and higher-order beliefs assign probability one to the lower-order beliefs of co-players. Strategy profile s is a Bayesian equilibrium if each player's strategy is a best reply to the belief hierarchy of this player's unique type. Therefore, the profile of trivial maps $\boldsymbol{\sigma} = s$ and associated belief hierarchies is an equilibrium of the trivial Bayesian game Γ if and only if it is an equilibrium of G according to the complete-information definition.

However, Harsanyi's methodology allows for a more flexible approach to complete-information equilibrium: A type structure in the sense of Harsanyi may have so-called "redundant" types, i.e., distinct types that nonetheless feature the same utility parameter and the same hierarchy of beliefs about utility parameters. This means that a *complete-information* game G can be equipped with a type structure $\mathcal{T}$ with *multiple redundant types*, thus obtaining a non-trivial Bayesian game $\Gamma = (G, \mathcal{T})$. Even though all types in Γ have the same utility function and hierarchical beliefs about utility functions, now an equilibrium profile $\boldsymbol{\sigma}$ can map different types to different strategies. As a consequence, in a Bayesian equilibrium players may be uncertain about the strategies and hierarchical beliefs about strategies of the co-players.[15] In other words, even if the utility functions are common knowledge and hence exogenous hierarchies of beliefs are trivially unique, there may be multiple hierarchies of endogenous beliefs. If there are sequential moves, this implies that players can change their mind about the co-players' intentions as the play unfolds on an equilibrium path, which is impossible according to the complete-information equilibrium concepts of Geanakoplos *et al.* (1989) and Battigalli & Dufwenberg (2009).[16]

[15] With standard (i.e., belief-independent) preferences, this is equivalent to the subjective correlated equilibrium concept (Brandenburger & Dekel 1987).

[16] See the notion of "polymorphic sequential equilibrium" of Battigalli *et al.* (2014).

3.8 Actual distribution of types and predictions

We adapt to psychological games the general notion of Bayesian equilibrium of Harsanyi (1967-68), which is inherently *subjective*, because players' beliefs are not derived from an objective distribution over types. In fact, our models do not assume that players know the objective statistical distribution of types and derive their exogenous beliefs from such distribution.[17] If this were the case, different types could have different beliefs only if the types of A and B were correlated, as in the case – for example – with assortative matching. But, under the random-matching structure typical of lab experiments, the types of A and B are objectively independent, with marginal probabilities given by the frequency distribution of types in the population from which subjects are drawn at random. Hence, conditioning on one's own type, a player cannot learn anything about the type of the co-player. With this, in a Bayesian equilibrium, a player's first- and second-order beliefs about the co-player would be type-independent, contrary to the findings of the experimental literature, which suggests instead that such beliefs are very heterogeneous.[18]

We analyze equilibria of Bayesian games with subjective beliefs to allow for such heterogeneity. Hence, our equilibrium analysis does not have to posit an objective statistical distribution on the type space. But, of course, such a distribution is necessary to derive statistical predictions. This is apparent in our leading example: The relative frequency of the trusting strategy I and of "optimistic" first-order beliefs ($\alpha_A > 1/2$) is just the fraction of A-agents in the population who believe that the high guilt type θ^H is more likely than the low guilt type θ^L. The relative frequency of the sharing strategy S coincides with the fraction of B-agents whose guilt type is θ^H. Finally, the relative frequency of "optimistic" initial second-order beliefs ($\beta_B^{\varnothing} > 1/2$) is the fraction of B-agents who believe that with more than 50% probability A deems θ^H more likely than θ^L.

A subtler question is whether we should expect to observe a positive correlation between (conditional) second-order beliefs and the propensity to Share. It turns out that this is the case if the guilt and epistemic components of B's type are statistically independent.[19] We extensively comment on statistical predictions in Section 6.

4 Role-dependent guilt

We analyze a model that generalizes the example of Section 3.3: Player A is commonly known to be a material payoff maximizer, but there are infinitely many possible guilt types of player B. If the agents playing in role A and B are drawn at random from the same population, as in most experiments, then we are assuming that the "potential" guilt sensitivity of an agent becomes an actual tendency to live up to the other's expectations only if this agent plays in the role of the "trustee" B. See the discussion in the Introduction.

4.1 Type structure

Since player A is commonly known to be selfish, $\Theta_A = \{0\}$. The set of possible guilt types of B is a closed interval $\Theta_B = [\theta^L, \theta^H]$ with $0 \leq \theta^L < 1$ and $\theta^H > 2$. The beliefs of each player i about the type of the co-player $-i$ are solely determined by the epistemic type $e_i \in \mathcal{E}_i$, with the convenient parametrization $\mathcal{E}_i = [0, 1]$. Therefore, $T_A = \{0\} \times \mathcal{E}_A \cong \mathcal{E}_A$ and $T_B = [\theta^L, \theta^H] \times [0, 1]$.

Each type of player A believes that the guilt and epistemic type of B are statistically independent. We let $e_A = t_A \in T_A$ parametrize A's subjective distribution over the guilt types of B: $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B < z] = G_{t_A}(z)$, where the cdf $G_{t_A} : \mathbb{R} \to [0, 1]$ has *support* $[\theta^L, \theta^H]$ and is *continuous* on

[17] In other words, we do not assume an "objective" common prior on the state space.

[18] See the references cited in the Introduction.

[19] This is just a sufficient condition for all models based on the Trust Minigame with guilt aversion.

(θ^L, θ^H), and $G_{t_A}(z)$ is *continuous* in t_A for each $z \in (\theta^L, \theta^H)$.[20] The beliefs about $\boldsymbol{\vartheta}_B$ of higher types first-order stochastically dominate those of lower types:

$$t'_A < t''_A \Rightarrow G_{t'_A}(z) > G_{t''_A}(z), \tag{7}$$

for all $z \in [\theta^L, \theta^H)$ and t'_A, $t''_A \in [0,1]$.

We also assume that there exist thresholds $\underline{t}_A > 0$ and $\bar{t}_A < 1$ such that every type $t_A < \underline{t}_A$ believes that $\boldsymbol{\vartheta}_B < 1$ with more than 50% probability, and every type $t_A > \bar{t}_A$ believes that $\boldsymbol{\vartheta}_B > 2$ with more than 50% probability:

$$\begin{array}{l} \underline{t}_A := \sup\left\{t_A : G_{t_A}(1) > \frac{1}{2}\right\} > 0, \\ \bar{t}_A := \min\left\{t_A : G_{t_A}(2) \leq \frac{1}{2}\right\} < 1. \end{array} \tag{8}$$

By continuity of $G_{t_A}(z)$ in t_A and z, and the stochastic-order assumption (7), $\underline{t}_A$ (resp., $\bar{t}_A$) is the unique solution to $G_{t_A}(1) = 1/2$ (resp., $G_{t_A}(2) = 1/2$), and $\underline{t}_A < \bar{t}_A$.

The marginal beliefs of each type t_A about the epistemic type of B are given by the same *continuous* cdf $F : \mathbb{R} \to [0,1]$ with *support* $[0,1]$.[21] The resulting belief function $\boldsymbol{\tau}_A : T_A \to \Delta(T_B)$ satisfies

$$\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B \leq z \cap \mathbf{e}_B \leq y] = G_{t_A}(z)F(y) \tag{9}$$

for all $t_A, y \in [0,1]$ and $z \in [\theta^L, \theta^H]$.

Each epistemic type e_B of B has beliefs about A's type given by a *continuous* cdf $F_{e_B} : \mathbb{R} \to [0,1]$ with *support* $[0,1]$. The resulting belief function $\boldsymbol{\tau}_B : T_B \to \Delta(T_A)$ satisfies

$$\boldsymbol{\tau}_B(\theta_B, e_B)[\mathbf{t}_A \leq x] = F_{e_B}(x), \tag{10}$$

for all $\theta_B \in [\theta^L, \theta^H]$ and $e_B, x \in [0,1]$.[22] 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:

$$e'_B < e''_B \Rightarrow \frac{1}{1 - F_{e'_B}(x)} \int_x^1 t_A \mathrm{d}F_{e'_B}(t_A) < \frac{1}{1 - F_{e''_B}(x)} \int_x^1 t_A \mathrm{d}F_{e''_B}(t_A) \tag{11}$$

for all $e'_B, e''_B \in [0,1]$ and $x \in [0,1)$. Intuitively, this means that higher epistemic types of B have higher beliefs about the (epistemic) type of A.[23]

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, $\boldsymbol{\alpha}_B(\theta_B, e_B)$, $\boldsymbol{\beta}_B^{\varnothing}(\theta_B, e_B)$, $\boldsymbol{\beta}_B^{I}(\theta_B, e_B)$, as we did in the leading example.

---

[20] G_{t_A} has support $[\theta^L, \theta^H]$ if it is strictly increasing on $[\theta^L, \theta^H]$, $G_{t_A}(z) = 0$ for $z < \theta^L$, and $G_{t_A}(z) = 1$ for $z \geq \theta^H$. We allow for atoms at θ^L and θ^H, as in the parametrized cdf

$$G_{t_A}(z) = \begin{cases} 0, & z < \theta^L, \\ 1 - t_A - \varepsilon + \varepsilon \frac{z - \theta^L}{\theta^H - \theta^L}, & \theta^L \leq z < \theta^H, \\ 1, & z \geq \theta^H, \end{cases}$$

which essentially gives back the leading example for ε small: $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B = \theta^H] = t_A$, $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B = \theta^L] = 1 - t_A - \varepsilon$.

[21] That is, F is strictly increasing on $[0,1]$ with $F(0) = 1 - F(1) = 0$.

[22] Conditions (9)-(10) imply that there is no perception of false consensus. See the discussion in Section 6.

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

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

for all $e'_B, e''_B \in [0,1]$ and $t_A \in [0,1)$. See Shaked & Shantikumar (2007, pp. 16-17). Notice that this stochastic-order property holds, but only weakly, in the leading example of Section 3.3.

4.2 Equilibrium analysis

By Remark 1, there is a pooling equilibrium with no trust and no cooperation. Henceforth, we study the non-degenerate equilibria, that is, those where a positive fraction of A-types choose In. Our assumptions imply that all non-degenerate equilibria exhibit threshold decision functions whereby higher types choose the pro-social action (the threshold types cannot be atoms, therefore their choices are immaterial). We call such functions "monotone":

Definition 2 *Decision function $\boldsymbol{\sigma}_A$ is **monotone (increasing)** if there is a threshold $\hat{t}_A \in [0,1]$ such that, for every t_A,*

$$\begin{aligned} t_A &< \hat{t}_A \Rightarrow \boldsymbol{\sigma}_A(t_A) = O, \\ t_A &> \hat{t}_A \Rightarrow \boldsymbol{\sigma}_A(t_A) = I. \end{aligned}$$

*Decision function $\boldsymbol{\sigma}_B$ is **monotone (increasing) in** θ_B if, for every $e_B \in [0,1]$, there is a threshold $\hat{\theta}_B(e_B) \in [\theta^L, \theta^H]$ such that, for every θ_B,*

$$\begin{aligned} \theta_B &< \hat{\theta}_B(e_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = K, \\ \theta_B &> \hat{\theta}_B(e_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = S, \end{aligned}$$

*and it is **monotone (increasing) in** e_B if, for every $\theta_B \in [\theta^L, \theta^H]$, there is a threshold $\hat{e}_B(\theta_B)$ such that, for every e_B,*

$$\begin{aligned} e_B &< \hat{e}_B(\theta_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = K, \\ e_B &> \hat{e}_B(\theta_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = S. \end{aligned}$$

A non-degenerate equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ determines the endogenous belief functions $\boldsymbol{\alpha}_A$, $\boldsymbol{\beta}_B^{\varnothing}$ and $\boldsymbol{\beta}_B^I$ as in eq. (4)-(5)-(6), so that $\boldsymbol{\sigma}_A^{-1}(I)$ 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 θ_B and e_B.

The incentive conditions give

$$\begin{aligned} \boldsymbol{\alpha}_A(t_A) &< \frac{1}{2} \Rightarrow \boldsymbol{\sigma}_A(t_A) = O, \\ \boldsymbol{\alpha}_A(t_A) &> \frac{1}{2} \Rightarrow \boldsymbol{\sigma}_A(t_A) = I, \end{aligned} \tag{12}$$

$$\begin{aligned} 2 &< 4 - 2\theta_B \boldsymbol{\beta}_B^I(e_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = K, \\ 2 &> 4 - 2\theta_B \boldsymbol{\beta}_B^I(e_B) \Rightarrow \boldsymbol{\sigma}_B(\theta_B, e_B) = S, \end{aligned} \tag{13}$$

for all $t_A \in T_A$, $(\theta_B, e_B) \in T_B$.

Proposition 3 *Every non-degenerate equilibrium of the model given by (7)-(11) has the following structure:*

(a) *$\boldsymbol{\sigma}_A$ is monotone with threshold $\hat{t}_A \in [\underline{t}_A, \bar{t}_A]$, which is the unique solution to equation*

$$\boldsymbol{\alpha}_A(\hat{t}_A) = \frac{1}{2},$$

where

$$\boldsymbol{\alpha}_A(t_A) = \int_{[0,1]} \left(1 - G_{t_A}\left(1/\boldsymbol{\beta}_B^I(e_B)\right)\right) \mathrm{d}F(e_B)$$

for all $t_A \in [0,1]$.

(b) *$\boldsymbol{\sigma}_B$ is monotone in θ_B with threshold function $\hat{\theta}(e_B) = 1/\boldsymbol{\beta}_B^I(e_B)$, and monotone in e_B with threshold function $\hat{e}(\theta_B) = (\boldsymbol{\beta}_B^I)^{-1}(1/\theta_B)$.*

(c) *The endogenous beliefs of B satisfy*

$$\begin{aligned}
\boldsymbol{\alpha}_B(e_B) &= 1 - F_{e_B}(\hat{t}_A) > 0, \\
\boldsymbol{\beta}_B^{\varnothing}(e_B) &= 1 - \int_{[0,1]} \int_{[0,1]} G_{t_A}\left(1/\boldsymbol{\beta}_B^I(x)\right) \mathrm{d}F(x) \mathrm{d}F_{e_B}(t_A), \\
\boldsymbol{\beta}_B^I(e_B) &= \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \mathbf{t}_A > \hat{t}_A] \geq \frac{1}{2}
\end{aligned}$$

for all $e_B \in [0,1]$.

Proposition 3 characterizes the structure of all non-degenerate equilibria of the model. Such equilibria have monotone decision functions. By condition (7), the higher the epistemic type of A, the higher A's belief that the guilt type of B is high. As higher guilt types of B have a higher propensity to Share, A's first-order belief is increasing in t_A; hence, according to the incentive condition (12), all epistemic types of A higher than $\hat{t}_A$ choose In.

By (11), higher epistemic types of B hold higher second-order beliefs on the epistemic type of A conditional on I. Since all epistemic types of A higher than $\hat{t}_A$ choose In, $\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \mathbf{t}_A > \hat{t}_A]$ is increasing in e_B. Hence, incentive condition (13) implies that the decision function of B is monotone both in θ_B and in e_B. Such decision function is characterized by a decreasing threshold function $\hat{e}(\theta_B)$, as shown in Figure 3.

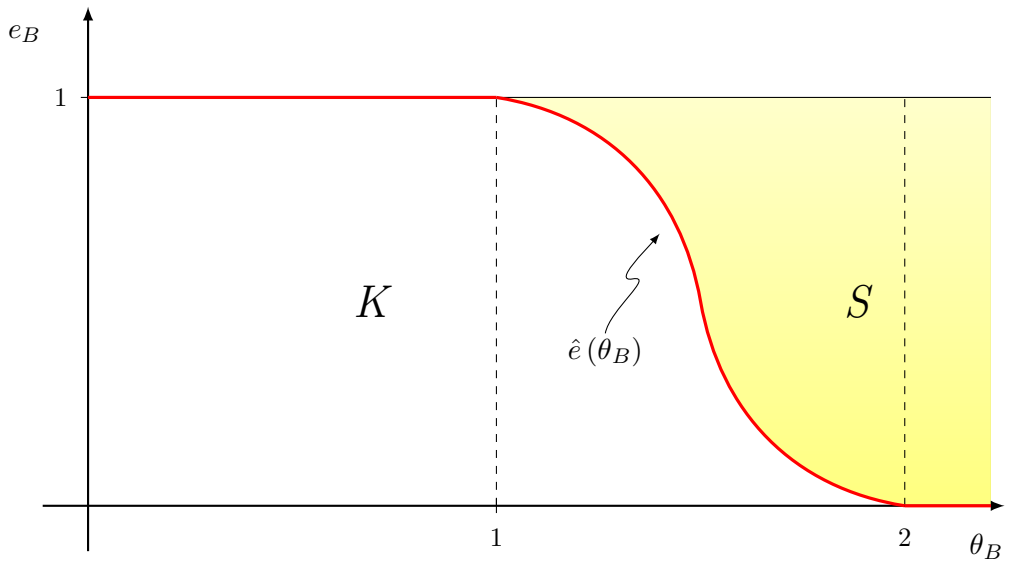

Figure 3: Equilibrium choice of B

As anticipated in Section 3.4, our assumptions imply that non-degenerate equilibria are the only ones consistent with forward-induction reasoning: Choice I signals that $\boldsymbol{\alpha}_A \geq 1/2$. Therefore, $\boldsymbol{\beta}_B^I(e_B) \geq 1/2$, and A predicts that all the types with $\theta_B > 2$ would Share; thus, $\boldsymbol{\alpha}_A(t_A) \geq \mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B > 2]$. By assumption, $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B > 2] > 1/2$ for all $t_A \in (\bar{t}_A, 1]$; hence, the measure of the set of $A-$types trusting B is at least $1 - \bar{t}_A > 0$.

On the other hand, $1 - \boldsymbol{\alpha}_A(t_A) \geq \mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B < 1]$. By assumption, $\mathbb{P}_{t_A}[\boldsymbol{\vartheta}_B < 1] < 1/2$ for all $t_A \in [0, \underline{t}_A)$; hence, the measure of the set of A-types that do not trust B is at least $\underline{t}_A > 0$. Therefore, there is heterogeneity of behavior and of endogenous beliefs among A-types. As for B, heterogeneity of behavior is quite obvious from incentive condition (13), given that B forward

inducts, and that there are types with $\theta_B > 2$ and types with $\theta_B < 1$. Heterogeneity of endogenous beliefs follows from forward-induction reasoning and our assumptions about exogenous beliefs.

5 Role-independent guilt

We now analyze a model of situations where the agents playing in role A and B are drawn from the same population, as in most games played in the lab. If guilt aversion is not affected by the role played in the game, the type structure must be symmetric. To anticipate, the main difference with the model of Section 4 is that here also player A may experience guilt feelings triggered by the expectation of B's disappointment. This is related to other differences between the assumed type structures.

Player B can only be disappointed after the terminal history O, in which case the extent of his disappointment also depends on what he plans to do in the subgame, i.e., his 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

$$\max\{0, \mathbb{E}_B[\mathbf{m}_B] - m_B(O)\} = \mathbb{E}_B[\mathbf{m}_B] - 1 = \begin{cases} \alpha_B, & \text{if } s_B = S, \\ 3\alpha_B, & \text{if } s_B = K. \end{cases} \tag{14}$$

We can represent this strategic situation with a psychological game parametrized by the guilt sensitivities θ_A and θ_B. To analyze such 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, 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 Figure 4.[24]

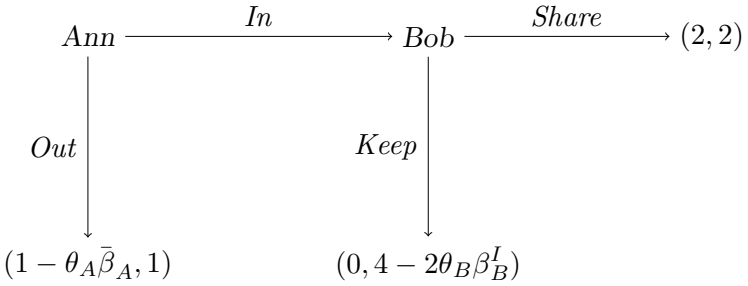

Figure 4: The Trust Minigame with psychological utilities of A and B

[24]Here we use an observation by Battigalli & Dufwenberg (2009): A psychological utility function of the form $u_i : \Theta_i \times Z \times \mathcal{H}_i \times \mathcal{H}_{-i} \to \mathbb{R}$ 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. As the example shows, this may require replacing low-order initial beliefs of others with own higher-order conditional beliefs.

5.1 Type structure

The possibility that both players can feel guilt complicates the analysis. Therefore we introduce some restrictive assumptions on the type space in order to maintain tractability of the model. In particular, while keeping the continuum of epistemic types on both sides, we now assume that each player i's guilt type can only take two values, low or high, i.e. $\theta_i \in \{\theta^L, \theta^H\}$, with $\theta^L = 0$ and $\theta^H > 1$. As in the previous section, we assume that $\mathcal{E}_i = [0,1]$, and that each epistemic type of each player i believes that the guilt and epistemic types of player $-i$ are independent.[25] Specifically, we model the exogenous beliefs of both players as we did for player A in the leading example of Section 3.3: e_i parametrizes i's subjective probability of the high-guilt type of the co-player: $\boldsymbol{\tau}_i(\theta_i, e_i)\left[\boldsymbol{\vartheta}_{-i} = \theta^H\right] = e_i$. This implies that, for each i, $t_i = (\theta_i, e_i) \in \{\theta^L, \theta^H\} \times [0,1] = T_i$ and that we can write $\boldsymbol{\tau}_i : [0,1] \to \Delta\left(\{\theta^L, \theta^H\} \times [0,1]\right)$. As a consequence, the endogenous second-order beliefs of players A and B are independent of their guilt type. 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 *support* $[0,1]$. Thus,

$$\forall e_i \in [0,1], \forall x,\ \boldsymbol{\tau}_i(e_i)[\boldsymbol{\vartheta}_{-i} = \theta^H \cap \mathbf{e}_{-i} \leq x] = e_i F(x), \tag{15}$$

where F is strictly increasing on $[0,1]$ and admits a density f. Note that here the epistemic type of B parametrizes an exogenous first-order belief, i.e. B's subjective probability that the guilt type of A is high. By contrast, in the model of Section 4, there is only one possible guilt type of A and e_B parametrizes the exogenous second-order belief of B.

By eq. (15), i's expectation of $\mathbf{e}_{-i}$ is independent of e_i, hence we write $\mathbb{E}_{e_i}[\mathbf{e}_{-i}] = \mathbb{E}[\mathbf{e}_{-i}]$.[26] To simplify the exposition and avoid tedious discussions of subcases in the equilibrium analysis, we assume that this expectation is not too low:

$$\mathbb{E}[\mathbf{e}_{-i}] > \frac{1}{3}. \tag{16}$$

5.2 Equilibrium analysis

Remark 1 applies to this model as well; hence, there is a pooling equilibrium with no trust and no cooperation. Henceforth, we focus on the characterization of equilibria $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ such that $\boldsymbol{\sigma}_A^{-1}(I)$ has positive measure, that is, the non-degenerate equilibria. The most important difference with the model of Section 4 is that, here, choice I is not a clear signal of A's trust, because A's guilt type may be high, and high-guilt type (θ^H, e_A) may choose I in equilibrium even if $\boldsymbol{\alpha}_A(e_A) < 1/2$, to avoid disappointing B. Specifically, eq. (14) shows that B's disappointment is maximal when he plans to Keep. Therefore, choice I may be interpreted as a signal that A's guilt type is high, A expects B to Keep, and goes In to prevent B's disappointment. The following result is a stark illustration of this phenomenon.

Proposition 4 *In the model given by (15)-(16) the following is an equilibrium:* $\boldsymbol{\sigma}_A(\theta^L, e_A) = O$, $\boldsymbol{\sigma}_A(\theta^H, e_A) = I$, *and* $\boldsymbol{\sigma}_B(\theta^L, e_B) = \boldsymbol{\sigma}_B(\theta^H, e_B) = K$ *for all* 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 all* e_A *and* e_B.

The equilibrium described by Proposition 4 is structurally different from the non-degenerate equilibria of Section 4: Here, every B-type Keeps and, despite this, high-guilt types of A choose I, because the prospective guilt from disappointing B prevails over the monetary incentive. The following proposition characterizes the other non-degenerate equilibria.

[25] In the model with role-dependent guilt this assumption holds trivially for the beliefs of B about A, because there is only one possible guilt type of A.

[26] All expectations not indexed by the epistemic type e_i are determined by the common marginal cdf F on $\mathcal{E}_{-i} = [0,1]$.

Proposition 5 *In the model given by (15)-(16) the non-degenerate equilibria* $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ *with heterogeneous behavior by B-types have the following structure:* $\boldsymbol{\sigma}_B(\theta^L, e_B) = K$ *for every* e_B*, the decision functions* $\boldsymbol{\sigma}_A(\theta^n, \cdot)$ *(*$n \in \{L, H\}$*) are monotone increasing,* $\boldsymbol{\sigma}_B(\theta^H, \cdot)$ *is monotone decreasing, and the corresponding 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* $0 \leq \hat{e}_B^H < 1$*, are such that*

(a)

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

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

and

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

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

for every e_A*, hence* $\boldsymbol{\alpha}_A(\cdot)$ *is increasing, and* $\bar{\boldsymbol{\beta}}_A(\cdot)$ *is decreasing;*

(b)

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

and

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

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

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

for every e_B*, hence* $\boldsymbol{\alpha}_B(\cdot)$ *is increasing,* $\boldsymbol{\beta}_B^{\varnothing}(\cdot)$ *is constant, and* $\boldsymbol{\beta}_B^I(\cdot)$ *is decreasing.*

The equilibria of Proposition 5 are more similar to those of Section 4. In both models, higher epistemic types of A are more confident that the guilt type of B is high and hence B would Share. As for player B, $\boldsymbol{\beta}_B^I(e_B)$ is decreasing in e_B, and so is the decision function $\boldsymbol{\sigma}_B(\theta^H, e_B)$. The reason why $\boldsymbol{\beta}_B^I(e_B)$ is decreasing, unlike the model of Section 4 where it is increasing, is that here e_B parametrizes the *first-order*, not the second-order beliefs of B.[27] A high (resp., low) epistemic type of B thinks that A's guilt type is likely to be high (resp. low). Therefore, high e_B's tend to explain choice I as the result of A's high guilt aversion despite the fact that α_A is low, whereas low e_B's think that A is selfish and explain I with a high α_A.

6 Discussion

We finally discuss the relevance of our models for experimental work, and then we offer our methodological perspective on the use of the Bayesian equilibrium concept.

[27] In Section 4 the exogenous first-order beliefs of B are trivial, because B knows that A is selfish.

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-degenerate 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.[28] Such predictions can be used to organize experimental data.

Heterogeneity and correlations 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 large fraction of subjects who exhibit intermediate beliefs.[29] Furthermore, if the epistemic component of players' types, $\mathbf{e}_i$, is statistically independent of the guilt component, $\boldsymbol{\vartheta}_i$, then we should observe positive correlation between pro-social actions and endogenous second-order beliefs, for player B and – in the case of role-independent guilt – also for player A. 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. Without relying on equilibrium analysis, Charness & Dufwenberg (2006) derive such positive correlation for B-subjects (supported by the data) from the assumption that, in our notation, $\boldsymbol{\vartheta}_B$ is statistically independent of the conditional second-order belief $\boldsymbol{\beta}_B^I$. This is consistent with our equilibrium analysis, but instead of *assuming* independence between $\boldsymbol{\vartheta}_B$ and the endogenous belief $\boldsymbol{\beta}_B^I$, we *derive* it from the independence between $\boldsymbol{\vartheta}_B$ and $\mathbf{e}_B$, which are both exogenous.

Statistical independence between the guilt and epistemic components 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. Adding such positive correlation to the model with role-independent guilt yields a negative correlation between the conditional second-order beliefs of B-subjects and 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 the conditional second-order belief. On the other hand, when A is known to be selfish (role-dependent guilt), we may have 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 components tends to strengthen the positive correlation between the pro-social action and the conditional second-order belief.[30]

[28] Recall that our models specify interactive beliefs for each type, but are silent on the actual distribution of types.

[29] Attanasi *et al.* (2013) instead implement in an experimental treatment a situation close to complete information, hence one where the support of exogenous beliefs for each matched pair of subjects is very small. They show that in this treatment endogenous beliefs tend to be extreme, as predicted by the complete information theory, whereas in the control treatment (incomplete information) they are indeed heterogeneous and mostly have intermediate values.

[30] 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' types, and that each player deems the epistemic component of the co-player type to be independent of the guilt component. However, 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

Further analysis of beliefs The theoretical insights of our models can be used by experimental economists to extend the elicitation and analysis of players' beliefs, design new experiments and explain previous experimental results.

First of all, our theoretical analysis highlights the importance of beliefs that are not considered in the experimental literature on the Trust Game, specifically, the *second-order* beliefs of A, and the *initial* beliefs of B.[31] The former are relevant when also A may be guilt averse: As discussed above, the presence of such guilt aversion should yield a positive correlation between A's second-order beliefs and the propensity to go In. New experiments eliciting such beliefs could check for such correlation, thus providing indirect evidence on the role dependence of guilt aversion. As for B's beliefs, choice should correlate with conditional beliefs, but also initial beliefs are relevant because they reflect B's strategic reasoning before playing the game.[32]

Our analysis also provides a potential explanation of why the conditional second-order beliefs of B-subjects do not conform to the classical forward-induction argument: If A is known to be selfish, then $\boldsymbol{\beta}_B^I = \mathbb{E}_B(\boldsymbol{\alpha}_A|I) \geq 1/2$ because A goes In only if $\boldsymbol{\alpha}_A \geq 1/2$. But, as shown in Section 5, if A is perceived by B as potentially guilt averse, action In may be interpreted as a desire not to disappoint B, hence it may well be the case that $\beta_B^I < 1/2$. Indeed, experimental data show that a significant fraction of B-subjects hold such low conditional second-order beliefs.[33]

6.2 Applicability of subjective Bayesian equilibrium

Our use of Bayesian equilibrium to model behavior and endogenous beliefs in games with belief-dependent preferences deserves discussion. Here we first explain why the traditional justification that equilibrium is attained through learning does not apply, then we elaborate on our interpretation of Bayesian equilibrium analysis.

Equilibrium and learning It is frequently argued that equilibrium analysis is appropriate to organize data because agents learn equilibrium behavior by playing a game many times against randomly matched co-players. But here we do not rely on such justification for several reasons.

First and most importantly, in so far as we aim at organizing experimental data, we must take into account that in most experiments on the Trust Game and other social dilemmas subjects play the game just once, or perhaps a few times, 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 may be different from a Nash or Bayesian equilibrium if agents have belief-dependent preferences.

A third, related issue is that we use the general, *subjective* notion of Bayesian equilibrium, because we assume that players do not know the objective distribution of types. Then, even with standard preferences, Bayesian equilibrium is not the right tool to capture self-confirming patterns of behavior. The reason is that Bayesian equilibrium postulates that players have correct conjectures about the true (type-dependent) decision functions of co-players, but this assumption can be justified by learning only in those rare situations where agents obtain sufficient information feedback to correctly identify the actual decision functions. However, such fine information

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.

[31] See, e.g., Charness & Dufwenberg (2006), Ellingsen *et al.* (2010), Chang *et al.* (2011). More precisely, Chang *et al.* (2011) elicit B's first-order beliefs, though they do not use them in the analysis.

[32] The connection between strategic reasoning and hierarchies of initial beliefs in dynamic games is clarifed by the literature on epistemic game theory. See Dekel & Siniscalchi (2015) and references therein.

[33] For example, in Charness & Dufwenberg (2006) – where the forward-induction threshold is 1/2 (7/10) in treatments with 5-5 (7-7) outside option – only 42% (19/45) of B-subjects in the control treatment with 5-5 outside option and only 31% (15/48) of B-subjects in the control treatment with 7-7 outside option have a conditional second-order belief above the forward-induction threshold.

feedback typically allows to also identify the distribution of types, which yields an *objective* Bayesian-Nash equilibrium (cf. Dekel *et al.* 2004). When instead such strong assumptions about information feedback do not hold and we model steady states of learning dynamics, beliefs about parameters are not exogenous, as in subjective Bayesian equilibrium analysis, because they have to be consistent with the long-run frequencies of observations implied – *via* information feedback – by the frequencies of parameter values and choices (e.g., observations of realized monetary payoffs).

To sum up, many experiments about social dilemmas are not designed so as to make subjects choose recurrently, hence there cannot be any equilibrating process through learning. But even if equilibrium analysis aims at organizing data about stabilized behavior in situations of recurrent interactions, subjective Bayesian equilibrium is not the appropriate tool, and a different approach is called for.

Equilibrium and strategic reasoning We instead use Bayesian equilibrium analysis to provide an orderly and consistent description of strategic reasoning in an incomplete information environment without assuming that behavior has stabilized, for example because subjects play just once. It has been shown that, if one drops – as we do – the assumption that exogenous beliefs are derived from an objective distribution, 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 (Battigalli & Siniscalchi 2003).[34] This result refers to equilibrium outcomes encompassing *all* the subjective Bayesian equilibrium models based on a given game form with parametrized utility functions. But, 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. In particular, our discussion of the model with role-dependent guilt emphasizes that some key results about non-degenerate equilibria follow from a forward-induction logic: All the guilt types $\theta_B > 2$ (those who would Share if they were sure that $\boldsymbol{\alpha}_A \geq 1/2$) do indeed Share given A's trusting action In, because they rationalize such action and infer that $\boldsymbol{\alpha}_A \geq 1/2$.[35] Predicting this, all the types of A who assign more than 50% probability to $\boldsymbol{\vartheta}_B > 2$ play the trusting action. On the other hand, the types who assign more than 50% probability to $\boldsymbol{\vartheta}_B < 1$ stay Out, because they understand that all the guilt types with $\theta_B < 1$ would Keep, independently of their beliefs. Given the heterogeneity of exogenous beliefs, forward induction is enough to imply heterogeneity of behavior and of endogenous beliefs.

Our equilibrium analysis goes beyond these key insights, yielding monotone decision functions and the correlations discussed in the previous subsection. Furthermore, in the model with role-independent guilt, the forward-induction logic does not have such clear-cut implications, because action In may be rationalized by a desire not to disappoint B rather than get a high monetary payoff.

It would be interesting to further depart from traditional equilibrium analysis and explore a rationalizability approach to guilt aversion in social dilemmas whereby some "natural" restrictions on beliefs are taken as given and commonly understood.[36] Battigalli *et al.* (2013) use this approach in the analysis of a cheap-talk sender-receiver game.

[34] This result relies on the existence of redundant types. Its earliest version is due to Brandenburger & Dekel (1987), who analyzed subjective correlated equilibria of games with complete infomation.

[35] As we did in Sections 2-5, we use "Share", "Keep", "Out" and "In" both as action labels and as words in the natural language

[36] See, e.g., Battigalli & Siniscalchi (2003) for such an approach to incomplete-information games with standard preferences.

Appendix

Proof of Proposition 3

The statement is implied by the following claims, which hold for every non-degenerate Bayesian equilibrium $(\boldsymbol{\sigma}_A, \boldsymbol{\sigma}_B)$ with endogenous belief functions $\boldsymbol{\alpha}_A$, $\boldsymbol{\alpha}_B$, $\boldsymbol{\beta}_B^{\varnothing}$ and $\boldsymbol{\beta}_B^I$.

Claim 6 *For all $e_B \in \mathcal{E}_B$, $\boldsymbol{\beta}_B^I(e_B) \geq \frac{1}{2}$.*

Proof In a non-degenerate equilibrium, $\boldsymbol{\sigma}_A(t_A) = I$ for a set of types with positive measure. By assumption, for each e_B, F_{e_B} has full support; hence $\mathbb{P}_{e_B}[\boldsymbol{\sigma}_A = I] > 0$ and $\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \boldsymbol{\sigma}_A = I]$ is well defined. Since $\boldsymbol{\sigma}_A(t_A) = I$ only if $\boldsymbol{\alpha}_A(t_A) \geq \frac{1}{2}$, then

$$\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \boldsymbol{\sigma}_A = I] \geq \mathbb{E}_{e_B}\left[\boldsymbol{\alpha}_A | \boldsymbol{\alpha}_A \geq \frac{1}{2}\right] \geq \frac{1}{2}.$$

□

Claim 7 *For every $(\theta_B, e_B) \in T_B$ with $\theta_B > 2$, $\boldsymbol{\sigma}_B(\theta_B, e_B) = S$.*

Proof By Claim 6, $\boldsymbol{\beta}_B^I(e_B) \geq \frac{1}{2}$; therefore, $\theta_B > 2$ implies $2 > 4 - 2\theta_B \boldsymbol{\beta}_B^I(e_B)$ and $\boldsymbol{\sigma}_B(\theta_B, e_B) = S$. □

Claim 8 *For every $t_A > \bar{t}_A$, $\boldsymbol{\sigma}_A(t_A) = I$.*

Proof By definition of $\bar{t}_A$, if $t_A > \bar{t}_A$ then $\boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B > 2] > 1/2$. By the stochastic-order assumption (7), it follows that

$$t_A > \bar{t}_A \Rightarrow \boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)[\boldsymbol{\sigma}_B = S] \geq \boldsymbol{\tau}_A(t_A)[\boldsymbol{\vartheta}_B > 2] > \frac{1}{2}.$$

Therefore, $\boldsymbol{\sigma}_A(t_A) = I$ for every $t_A > \bar{t}_A$. □

Claim 9 *Decision function $\boldsymbol{\sigma}_B$ is monotone in θ_B with threshold $\hat{\theta}(e_B) = 1/\boldsymbol{\beta}_B^I(e_B) \leq 2$.*

Proof Fix e_B arbitrarily. By incentive condition (13),

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

where the threshold $\hat{\theta}(e_B) = 1/\boldsymbol{\beta}_B^I(e_B)$. By Claim 6, $\hat{\theta}(e_B) \leq 2$. □

Claim 10 *Decision function $\boldsymbol{\sigma}_A$ is monotone with threshold $\hat{t}_A \in [\underline{t}_A, \bar{t}_A]$, which is the unique solution to equation*

$$\int_{[0,1]} G_{t_A}\left(1/\boldsymbol{\beta}_B^I(e_B)\right) \mathrm{d}F(e_B) = \frac{1}{2}. \tag{17}$$

Proof By Claim 9,

$$\boldsymbol{\alpha}_A(t_A) = \boldsymbol{\tau}_A(t_A)\left[\boldsymbol{\vartheta}_B > \frac{1}{\boldsymbol{\beta}_B^I(e_B)}\right] = \int_{[0,1]} \left(1 - G_{t_A}\left(1/\boldsymbol{\beta}_B^I(e_B)\right)\right) \mathrm{d}F(e_B).$$

Therefore, by the stochastic-order assumption (7), $t'_A < t''_A$ implies

$$\begin{aligned}\boldsymbol{\alpha}_A(t'_A) &= \int_{[0,1]} \left(1 - G_{t'_A}\left(1/\boldsymbol{\beta}_B^I(e_B)\right)\right) \mathrm{d}F(e_B) \\ &< \int_{[0,1]} \left(1 - G_{t''_A}\left(1/\boldsymbol{\beta}_B^I(e_B)\right)\right) \mathrm{d}F(e_B) = \boldsymbol{\alpha}_A(t''_A).\end{aligned}$$

Since $\boldsymbol{\alpha}_A(\underline{t}_A) \leq 1/2$ and $\boldsymbol{\alpha}_A(\bar{t}_A) \geq 1/2$, threshold $\hat{t}_A$ is the unique solution $\hat{t}_A \in [\underline{t}_A, \bar{t}_A]$ to eq. (17). Incentive condition (12) implies that $\boldsymbol{\sigma}_A$ is monotone with threshold $\hat{t}_A$. □

Claim 10 implies that $\boldsymbol{\beta}_B^{\varnothing}(e_B) = 1 - \int_{[0,1]} \int_{[0,1]} G_{t_A}\left(1/\boldsymbol{\beta}_B^I(x)\right) \mathrm{d}F(x) \mathrm{d}F_{e_B}(t_A)$.

Claim 11 *$\boldsymbol{\beta}_B^I$ is strictly increasing.*

Proof By Claim 10 and the assumption that each $F_{e_B}(\cdot)$ is continuous and strictly increasing on $[0,1]$,

$$\boldsymbol{\beta}_B^I(e_B) = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \boldsymbol{\sigma}_A = I] = \mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \mathbf{t}_A > \hat{t}_A],$$

where $\hat{t}_A$ is the threshold of Claim 10 and $\mathbb{E}_{e_B}[\boldsymbol{\alpha}_A | \mathbf{t}_A > \hat{t}_A]$ is strictly increasing in e_B by eq. (11). □

Claim 12 *Decision function $\boldsymbol{\sigma}_B$ is monotone in e_B with threshold function $\hat{e}(\theta_B) = (\boldsymbol{\beta}_B^I)^{-1}(1/\theta_B)$.*

Proof By Claims 9 and 11, $\boldsymbol{\beta}_B^I(e_B) = 1/\hat{\theta}_B(e_B)$ is increasing, hence invertible. Thus, incentive condition (13) implies that $\boldsymbol{\sigma}_B$ is monotone in e_B with threshold function $\hat{e}(\theta_B) = (\boldsymbol{\beta}_B^I)^{-1}(1/\theta_B)$. □

Claim 13 *The endogenous first-order belief of B is*

$$\boldsymbol{\alpha}_B(e_B) = 1 - F_{e_B}(\hat{t}_A) > 0.$$

Proof Claim 10 implies that

$$\boldsymbol{\alpha}_B(e_B) = 1 - F_{e_B}(\hat{t}_A).$$

The fact that $\hat{t}_A \leq \bar{t}_A$, together with the assumptions that $\bar{t}_A < 1$ and that each $F_{e_B}(\cdot)$ is continuous and strictly increasing on $[0,1]$ imply that $\boldsymbol{\alpha}_B(e_B) > 0$. □

■

Proof of Proposition 4

We must show that the candidate equilibrium where $\boldsymbol{\sigma}_A(\theta^L, e_A) = O$, $\boldsymbol{\sigma}_A(\theta^H, e_A) = I$, $\boldsymbol{\sigma}_B(\theta_B, e_B) = K$ for all e_A, θ_B, and e_B satisfies the incentive constraints. These decision functions imply that $\boldsymbol{\alpha}_A(e_A) = 0$ and $\boldsymbol{\alpha}_B(e_B) = \mathbb{P}_{e_B}\left[\boldsymbol{\vartheta}_A = \theta^H\right] = e_B$ for all e_A and e_B. Therefore,

$$\boldsymbol{\beta}_B^I(e_B) = \boldsymbol{\beta}_B^{\varnothing}(e_B) = 0$$

and

$$\bar{\boldsymbol{\beta}}_A(e_A) = \mathbb{E}_{e_A}\left[\mathbb{E}_{\mathbf{e}_B}[\mathbf{m}_B] - m_B(O)\right] = \mathbb{E}_{e_A}\left[1 - \boldsymbol{\alpha}_B + 4\boldsymbol{\alpha}_B - 1\right] = 3\mathbb{E}_{e_A}\left[\boldsymbol{\alpha}_B\right] = 3\mathbb{E}\left[\mathbf{e}_B\right] > 1,$$

where we used the definition of $\boldsymbol{\beta}_B^I$ and Bayes' rule, the definition of $\bar{\boldsymbol{\beta}}_A$, eq. (14), and condition (16).

The beliefs equations for A and the fact that $\theta^L = 0$ and $\theta^H > 1$ imply that incentive condition (12) holds for every type of A, whereas incentive condition (13) holds for every type of B because $\boldsymbol{\beta}_B^I = 0$.

■

Proof of Proposition 5

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 decision functions $\boldsymbol{\sigma}_A(\theta^n,\cdot):[0,1]\to\{I,O\}$, $\boldsymbol{\sigma}_B(\theta^n,\cdot):[0,1]\to\{S,K\}$, with $n=H,L$, and we show that $\boldsymbol{\sigma}_A(\theta^n,\cdot)$ is monotone increasing and $\boldsymbol{\sigma}_B(\theta^n,\cdot)$ is monotone decreasing. We also provide a characterization of some properties of the endogenous beliefs. We do so by proceeding through a series of claims.

The observation that low-guilt types of B choose K (Remark 14) is used to prove that $\boldsymbol{\sigma}_A(\theta^L,\cdot)$ is monotone (Claim 15). Then, we show that in every non-degenerate equilibrium A's expectation of B's disappointment, $\bar{\boldsymbol{\beta}}_A$, is strictly positive (Claim 16). Next, we prove that $\bar{\boldsymbol{\beta}}_A>0$ implies that B believes a high-guilt A to be strictly more likely to go In than a low-guilt A (Claim 17). We use this implication to characterize the monotonicity of the endogenous beliefs of B (Claims 18 and 19), and to prove that $\boldsymbol{\sigma}_B(\theta^H,\cdot)$ is monotone decreasing (Claim 19). Next, we prove that $\boldsymbol{\sigma}_A(\theta^H,\cdot)$ is monotone (Claim 20).

Remark 14, Claim 15, and Claims 18-20 taken together lead to Proposition 5.

Remark 14 *B's incentive condition, characterized by eq. (13) also in this model, implies that $\boldsymbol{\sigma}_B(\theta^L,e_B)=K$ for every e_B.*

Let μ_A be the common marginal belief of each type of A about the epistemic type of B, and let $E_B^{HS}=\left\{e_B:\boldsymbol{\sigma}_B(\theta^H,e_B)=S\right\}$. The next claim shows that both A's initial first-order belief and his θ^L-decision function are increasing in his epistemic type.

Claim 15 *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 Remark 14 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, $\boldsymbol{\tau}_A(e_A)[\boldsymbol{\vartheta}_B=\theta^H\cap\mathbf{e}_B\le y]=e_AF(y)$ 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 (12) for A when the guilt type is low 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-degenerate equilibrium, A necessarily expects to disappoint B by going Out. Formally:

Claim 16 *In every non-degenerate equilibrium, $\bar{\boldsymbol{\beta}}_A(e_A)>0$ for each e_A.*

Proof In a non-degenerate 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 $\mathcal{E}_A = [0,1]$ induced by cdf F, an exogenous marginal belief of player B. To ease notation, let

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

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 epistemic type $e_B \in (0,1)$,

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

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

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

□

Claim 17 *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 $\theta^H > 1$ by assumption, and $\bar{\boldsymbol{\beta}}_A(e_A) > 0$ by Claim 16. 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 $\mathcal{E}_A = [0,1]$ 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 $e_B \in (0,1)$,

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

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

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

Claim 18 *The endogenous first-order 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 endogenous first-order 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 17. □

Claim 19 *The endogenous second-order 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, $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ is monotone decreasing, that is

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

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 endogenous second-order belief of B 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 $\mathcal{E}_B = [0,1]$. 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 15 and 17, we obtain

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

whenever the conditional probabilities are well defined. 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 17 (note that $\mu_A[E_B^{HS}]$ may be zero). The incentive condition (13) 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 $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$ is monotone decreasing in e_B, and characterized by a threshold $\hat{e}_B^H$ that satisfies the incentive conditions stated in this claim. □

Claim 20 *The endogenous second-order belief of A is such that*

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

which is decreasing. Moreover $\boldsymbol{\sigma}_A\left(\theta^H, \cdot\right)$ is monotone (increasing), that is

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

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

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

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

$$\bar{\boldsymbol{\beta}}_A(e_A) = \mathbb{E}_{e_A}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K\right]\mathbb{P}_{e_A}\left[\boldsymbol{\sigma}_B = K\right] + \mathbb{E}_{e_A}\left[\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = S\right]\mathbb{P}_{e_A}\left[\boldsymbol{\sigma}_B = S\right].$$

Decomposing expected values and taking into account that $\mathbb{P}\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^L\right] = 0$, we obtain

$$\begin{aligned}\bar{\boldsymbol{\beta}}_A(e_A) &= \mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^L\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = K|\boldsymbol{\vartheta}_B = \theta^L\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^L\right] \\ &\quad +\mathbb{E}\left[3\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = K \cap \boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = K|\boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^H\right] \\ &\quad +\mathbb{E}\left[\boldsymbol{\alpha}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}\left[\boldsymbol{\sigma}_B = S|\boldsymbol{\vartheta}_B = \theta^H\right]\mathbb{P}_{e_A}\left[\boldsymbol{\vartheta}_B = \theta^H\right].\end{aligned}$$

Replacing probabilities with their specific expressions and using Claim 18, we obtain

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

Now observe that, since the random variables $\mathbf{e}_B$ and $\boldsymbol{\vartheta}_B$ are independent, we can write

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

Regrouping terms in the expression of $\bar{\boldsymbol{\beta}}_A(e_A)$, this simplifies to

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

Note that $\bar{\boldsymbol{\beta}}_A(\cdot)$ is decreasing:

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

This completes the proof of the first part of the claim.

To show that $\boldsymbol{\sigma}_A\left(\theta^H, \cdot\right)$ is monotone, we consider A's incentive condition, that is, type (θ^H, e_A) of A chooses I when

$$2\boldsymbol{\alpha}_A + \theta^H \bar{\boldsymbol{\beta}}_A(e_A) > 1.$$

Recall that Claim 15 shows that $\boldsymbol{\alpha}_A = \mu_A[E_B^{HS}]e_A$; hence, the incentive condition can be rewritten as

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

Next, we show that either (i) the left-hand side (LHS) is increasing in e_A, hence $\boldsymbol{\sigma}_A(\theta^H, \cdot)$ is monotone (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 $\partial\bar{\boldsymbol{\beta}}_A/\partial e_A$, we obtain:

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

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

$$\theta^H\left(\mu_B[E_A^{LI}] + \left(\mu_B[E_A^{HI}] - \mu_B[E_A^{LI}]\right)\mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]\right) > 1. \tag{18}$$

Note that, by Claim 19, $\mathbb{E}\left[\mathbf{e}_B\right] \geq \mathbb{E}[\mathbf{e}_B|\mathbf{e}_B < \hat{e}_B^H] = \mathbb{E}\left[\mathbf{e}_B|\boldsymbol{\sigma}_B = S \cap \boldsymbol{\vartheta}_B = \theta^H\right]$; therefore, we obtain the following inequalities, which imply that the LHS is larger than 1:

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

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

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

The second inequality holds by eq. (18) 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 conditions

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

□

Note that, given that the decision functions $\boldsymbol{\sigma}_B\left(\theta^H, \cdot\right)$, $\boldsymbol{\sigma}_A\left(\theta^L, \cdot\right)$ and $\boldsymbol{\sigma}_A\left(\theta^H, \cdot\right)$ are monotone and described respectively by thresholds $\hat{e}_B^H$, $\hat{e}_A^L$ and $\hat{e}_A^H$, we have that

$$\begin{aligned}
\mu_A[E_B^{HS}] \ &= \ F(\hat{e}_B^H), \\
\mu_B[E_A^{LI}] \ &= \ 1 - F(\hat{e}_A^L), \\
\mu_B[E_A^{HI}] \ &= \ 1 - F(\hat{e}_A^H).
\end{aligned}$$

This, together with Remark 14, Claim 15, and Claims 18-20 delivers the result stated in Proposition 5.

■

References

[1] Attanasi G., P. Battigalli and R. Nagel (2013). "Disclosure of Belief-Dependent Preferences in the Trust Game," IGIER Working Paper 506, Bocconi University.

[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*, 204-232. Houndmills: Palgrave McMillan.

[3] Barrett L.F. (2006). "Solving the Emotion Paradox: Categorization and the Experience of Emotion," *Personality and Social Psychology Review*, 10, 20-46.

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

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

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

[7] Battigalli P., G. Charness and M. Dufwenberg (2013). "Deception: The Role of Guilt," *Journal of Economic Behavior and Organization*, 93, 227-232.

[8] Battigalli P., M. Dufwenberg and A. Smith (2014). "Frustration and Anger in Games," mimeo, Bocconi University.

[9] 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.

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

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

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

[13] Buskens V. and W. Raub (2013). "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust," in R. Wittek, T.A.B. Snijders and V. Nee (Eds.), *Handbook of Rational Choice Social Research*, 113-150. Stanford: Stanford University Press.

[14] Caplin A. and J. Leahy (2004). "The Supply of Information by a Concerned Expert," *Economic Journal*, 114, 487-505.

[15] Carter, C.S. and E.B. Keverne (2002). "The Neurobiology of Social Affiliation and Pair Bonding," in D. Pfaff, A.P. Arnold, A.M. Etgen, S.E. Fahrbach and R.T. Rubin (Eds.), *Hormones, Brain, and Behavior*, 299-337. San Diego: Academic Press.

[16] 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.

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

[18] Charness, G. and M. Dufwenberg (2011). "Participation," *American Economic Review*, 101, 1213-1239.

[19] 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,* Vol. 2, (forthcoming). Princeton: Princeton University Press.

[20] Dekel E., and M. Siniscalchi (2015): "Epistemic Game Theory," in P. Young and S. Zamir (Eds.), *Handbook of Game Theory*, Vol. 4, 619-702. Amsterdam: North Holland (Elsevier).

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

[22] Dufwenberg M. (2002). "Marital Investment, Time Consistency and Emotions," *Journal of Economic Behavior and Organization*, 48, 57-69.

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

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

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

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

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

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

[29] 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.

[30] 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), 21-40. New York: Psychology Press.

[31] Kosfeld M., M. Heinrichs, P. Zak, U. Fischbacher and E. Fehr (2005). "Oxytocin Increases Trust in Humans," *Nature*, 435, 673-676.

[32] Ong D. (2011). "Fishy Gifts: Bribing with Shame and Guilt," SSRN Working Paper 1303051.

[33] 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.

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

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

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

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

[38] Zak P.J. (2008). "The Neurobiology of Trust," *Scientific American*, 298, 88-95.

[39] Zak P.J., R. Kurzban and W.T. Matzner (2005). "Oxytocin is Associated with Human Trustworthiness," *Hormones and Behavior*, 48, 522-527.