Disclosure of Belief–Dependent Preferences in a Trust Game

Giuseppe Attanasi (Sapienza University of Rome)
Pierpaolo Battigalli (Bocconi University and IGIER, Milan)
Elena Manzoni (University of Bergamo)
Rosemarie Nagel (ICREA, Universitat Pompeu Fabra, Barcelona GSE)

Online Appendix B

In this appendix, we provide:

B.1 the derivation of B's hypothetical payback function $\xi(\alpha)$, a characterization of how its slope depends on guilt and reciprocity, and a graphical intuition for its different possible quasi-convex shapes;

B.2 the complete derivation of the theoretical predictions for the Trust Minigame (rationalizability under complete and incomplete information, and Bayesian equilibrium under incomplete information).

B.1 Hypothetical payback function: derivation and shapes

Here we provide the details of the derivation of B's payback function as a best response to the hypothetical questions in Table 3 (phase 2 questionnaire). Our baseline assumption is that B fills in the payback scheme of Table 3 as if the amount x that he hypothetically gives back to A were really given to A, thus implementing the distribution $(m_A, m_B) = (x, 4-x)$ with $x \in [0,4]$. The expected payoff for A of action *Continue* is 2α, hence, modeling disappointment as in Battigalli & Dufwenberg (2007), $D_A(\alpha, x) = \max\{0, 2\alpha - x\}$, where α is the subjective probability assigned by A to *Share*.

Recall that we assume that B's preferences are described by the following utility function:

$$u_i(m_i, m_j, \alpha_j) = \ln(1+m_i) - \frac{G_i}{4} \cdot [D_j(\alpha_j, m_j)]^2 + R_i \cdot K_j(\alpha_j) \cdot m_j, \tag{1}$$

where G_i and R_i respectively parametrize sensitivity to guilt and reciprocity. The kindness of action *Continue* as a function of α is modeled as in Dufwenberg & Kirchsteiger (2004), which implies that *Continue* is always a kind action, but less so the more A expects B to share (the higher α). Indeed, the higher α, the lower the increase in B's payoff that A expects to induce by choosing *Continue* rather than *Dissolve*. Specifically, the equitable payoff of B in A's eyes is the average of B's expected payoff under *Continue* and *Dissolve*: $m_B^e(\alpha) = \frac{1}{2}[\mathbb{E}_A(\widetilde{m}_B; Diss, \alpha) + \mathbb{E}_A(\widetilde{m}_B; Cont, \alpha)] = \frac{1+(4-2\alpha)}{2} = \frac{5}{2} - \alpha$; hence, the kindness of *Continue* is $K_A(\alpha) = (4-2\alpha) - \left(\frac{5}{2} - \alpha\right) = \frac{3}{2} - \alpha$.

Plugging $D_A(\alpha, x)$ and $K_A(\alpha)$ in (1), we obtain the maximization problem

$$\max_{x\in[0,4]} \left\{ \ln(5-x) - \frac{G}{4} \cdot [\max\{0, 2\alpha - x\}]^2 + R \cdot \left(\frac{3}{2} - \alpha\right) \cdot x \right\}. \tag{2}$$

However, there is a possible confound. Since we put the B responder in a hypothetical situation in which he has "transgressed,"[1] we have to allow for the possibility that B chooses a higher x than implied by the solution to (2). This is because the transgression puts him in an *ex-post* negative affective state that can be alleviated by giving more than he would *ex ante*. Such "moral cleansing" (Sachdeva *et al.* 2009) is consistent with experimental findings by psychologists and economists (Ketelaar & Au 2003, Silfver 2007, and Brañas-Garza *et al.* 2013).[2] Indeed, several B-subjects in our experiment provided comments to the filled-in

[1] Each B-subject in phase 2 is asked to consider the following hypothetical situation: "suppose that [...] A chose *Continue* and you chose *Take*, hence you got €4 and left A with €0 in his/her pocket." See the experimental instructions in *Online Appendix A*.

[2] In particular, Silfver (2007) shows that the action-tendency associated to guilt is to engage in "repair behavior." Note that, instead, the theory of guilt aversion (Dufwenberg 2002, Battigalli & Dufwenberg 2009) highlights avoidance of the anticipated negative valence associated with guilt.

questionnaire in Table 2 that are in line with such repair-behavior hypothesis.[3] Therefore, we introduce in the maximization problem an ex-*p*ost feeling-mitigation parameter $p \in [0,1]$ that boosts the payback x by adding to α in the disappointment function and subtracting from it in the kindness function. The modified maximization problem is

$$\max_{x\in[0,4]} \left\{ \ln(5-x) - \frac{G}{4} \cdot [\max\{0, 2(\alpha+p)-x\}]^2 + R \cdot \left(\frac{3}{2} - (\alpha - p) \right) \cdot x \right\}. \qquad (3)$$

This functional form has the advantage of being strictly concave in the payback x, with a decreasing, continuous derivative

$$U_x = -\frac{1}{5-x} + \frac{G}{2} \cdot \max\{0, 2p+2\alpha-x\} + R\left(\frac{3}{2} + p - \alpha\right).$$

By strict concavity, (3) has a unique solution $x^* = \xi(\alpha)$. We call $\xi(\alpha)$ the **payback function.**

Let us now describe the main features of the payback function $\xi(\alpha)$ and its dependence on guilt, reciprocity, and ex-post feeling-mitigation components. Proposition B.1.1 shows how the slope of the payback function $\xi(\alpha)$ depends on the comparison between guilt and reciprocity components. In each case, $\xi(\alpha)$ is quasi-convex, that is, either monotone or U-shaped.

Proposition B.1.1 *Consider the range of α where an interior solution obtains (i.e., $G(p+\alpha) + R(3/2+p-\alpha) > 1/5$, $R(3/2+p-\alpha) < 1$). The payback function $\xi(\alpha)$ is*
(i) increasing *if $G > R$ and $R \leq \underline{R}(p)$,*
*(ii) first decreasing and then increasing (*U-shaped*) if $G > R$ and $\underline{R}(p) < R < \overline{R}(p)$,*
(iii) decreasing *if either $G < R$ or $R \geq \overline{R}(p)$,*
(iv) constant *if $G = R$ and $R \leq \underline{R}(p)$,*
where $\underline{R}(p) = 1/[(5-2p)(3/2+p)]$ and $\overline{R}(p) = 1/[(3-2p)(1/2+p)]$. Furthermore, $\xi(\alpha)$ is increasing in a neighborhood of α only if $\xi(\alpha) < 2p + 2\alpha$.

To prove Proposition B.1.1 we therefore need to obtain B's best-response function to A's initial belief about his strategy *Share*:

$$\xi(\alpha; G, R, p) := \arg\max_{x\in[0,4]} U(x; \alpha, G, R, p),$$

both in the case B's payback depends on A's disappointment and in the case it does not.

[3] In *Online Appendix C* we report the answers to the debriefing questions about subjects' interpretation of their filled-in questionnaire: (a) "Explain the meaning of the values you entered in the Hypothetical Payback Scheme. Did you enter these values according to a specific feeling?" and (b) "What kind of relationship is there between this feeling and your partner's guess about you choosing Share?"

Let us first consider the case where A's disappointment matters to B, because he gives back to her less than (the sum of his ex-post feeling-mitigation payoff and) what A would have expected to get after *Continue*, i.e. $x \in [0, 2p + 2\alpha)$. In this case, $U_x(x; \alpha, G, R, p) = 0$ yields:

$$-\frac{1}{5-x} + \frac{G}{2}(2p + 2\alpha - x) + R\left(\frac{3}{2} + p - \alpha\right) = 0$$

which, given that $x < 5$ by construction, can be rewritten as

$$\frac{G}{2}x^2 - \left[G\left(\frac{5}{2} + p + \alpha\right) + R\left(\frac{3}{2} + p - \alpha\right)\right]x + 5\left[G(p+\alpha) + R\left(\frac{3}{2} + p - \alpha\right)\right] - 1 = 0.$$

The determinant of the second-order equation in x is

$$\Delta = \left[G\left(\frac{5}{2} - p - \alpha\right) - R\left(\frac{3}{2} + p - \alpha\right)\right]^2 + 2G,$$

which is positive for $G > 0$. Therefore, the two roots of the second-order equation in x are

$$x_{1/2} = \frac{G\left(\frac{5}{2} + p + \alpha\right) + R\left(\frac{3}{2} + p - \alpha\right) \pm \sqrt{\left[G\left(\frac{5}{2} - p - \alpha\right) - R\left(\frac{3}{2} + p - \alpha\right)\right]^2 + 2G}}{G}.$$

Given that the greater of the two roots is never lower than $2p+2\alpha$ for any nonnegative vector (α, G, R, p), the only acceptable solution is the smaller root. Therefore, the best-response function when A's disappointment matters, $\xi_D(\alpha; G, R, p)$, is:

$$\xi_D(\alpha; G, R, p) = \left(\frac{5}{2} + p + \alpha\right) + \frac{R}{G}\left(\frac{3}{2} + p - \alpha\right) - \sqrt{\left[\left(\frac{5}{2} - p - \alpha\right) - \frac{R}{G}\left(\frac{3}{2} + p - \alpha\right)\right]^2 + \frac{2}{G}}$$

for $R\left(\frac{3}{2} + p - \alpha\right) \in \left(\frac{1}{5} - G\left(p + \alpha\right), \frac{1}{5-2p-2\alpha}\right]$. If $R\left(\frac{3}{2} + p - \alpha\right) \in \left[0, \frac{1}{5} - G\left(p + \alpha\right)\right]$, then $\xi_D(\alpha; G, R, p) = 0$.

Let us now consider the case where A's disappointment does not matter to B, because $x \in [2p + 2\alpha, 4]$. In this case, $U_x(x; \alpha, G, R, p) = 0$ yields:

$$-\frac{1}{5-x} + R\left(\frac{3}{2} + p - \alpha\right) = 0.$$

Thus, the best-response function when A's disappointment does not matter is:

$$\xi_{NoD}(\alpha; R, p) = 5 - \frac{2}{R(3 + 2p - 2\alpha)}$$

for $R\left(\frac{3}{2}+p-\alpha\right) \in \left(\frac{1}{5-2p-2\alpha}, 1\right]$. For $R\left(\frac{3}{2}+p-\alpha\right) \in (1, +\infty)$ it is $\xi_{NoD}(\alpha; R, p) = 4$.

To sum up, B's payback function is given by the following formula:

$$\xi(\alpha; G, R, p) = \begin{cases} 0 & \text{if} \quad R\left(\frac{3}{2}+p-\alpha\right) \in \left[0, \frac{1}{5} - G\left(p+\alpha\right)\right], \\ \xi_D(\alpha; G, R, p) & \text{if} \quad R\left(\frac{3}{2}+p-\alpha\right) \in \left(\frac{1}{5} - G\left(p+\alpha\right), \frac{1}{5-2p-2\alpha}\right], \\ \xi_{NoD}(\alpha; R, p) & \text{if} \quad R\left(\frac{3}{2}+p-\alpha\right) \in \left(\frac{1}{5-2p-2\alpha}, 1\right], \\ 4 & \text{if} \quad R\left(\frac{3}{2}+p-\alpha\right) \in (1, +\infty). \end{cases}$$

Now we provide a graphical intuition for the quasi-convex shapes of $\xi(\alpha)$ according to Proposition 1.

The first-order condition for an interior solution of the modified maximization problem (3) can be better understood in terms of the "marginal cost" and "marginal benefit" of the payback x. The first-order condition can be rewritten as:

$$MC(x) \equiv \frac{1}{5-x} = \frac{G}{2} \cdot \max\{0, 2p + 2\alpha - x\} + R \cdot \left(\frac{3}{2} + p - \alpha\right) \equiv MB(x). \tag{4}$$

This helps us understand how the payback changes as a function of the first-order belief α and of parameter shifts. In Figure B.1 we draw the MC and MB schedules under different cases (Figures B.1-a,b) and trace how their intersection is affected by parameter shifts (Figures B.1-c,d). Figure B.1-a shows a typical solution when $G > R$ and α is high. In this case, an increase in α *increases* the payback (Figure B.1-c). Figure B.1-b shows a typical solution when $R > G$ and α is low. In this case, an increase in α *decreases* the payback (Figure B.1-d).

The following discussion of Figure B.1 provides an intuitive proof of the four results (four possible shapes of the payback function) in Proposition 1.

First note that an interior solution obtains if $\max_{x\in[0,4]}\{MB(x) - MC(x)\} > 0$ and $\min_{x\in[0,4]}\{MB(x) - MC(x)\} < 0$. This gives the condition on the range of α. The payback function is *increasing* if and only if the part of the MB schedule with negative slope shifts upward with an increase in α, and the flat part of the MB schedule is always below the MC schedule (see Figures B.1-a, B.1-c). The first condition holds if and only if $G > R$; the second condition holds if and only if $R(3/2 + p - \alpha) \leq MC(2p + 2\alpha) \equiv 1/(5 - 2p - 2\alpha)$ for every $\alpha \in [0, 1]$, that is, if and only if $R \leq 1/[(5 - 2p)(3/2 + p)]$: the reciprocity component is low enough that A's disappointment matters to B. This explains result (i); the intuition for result (iv) (*constant* payback function) is similar.

The payback function is *decreasing* if either the part of MB with negative slope shifts

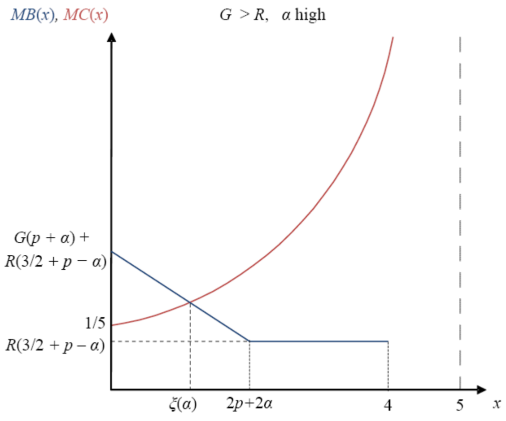

Fig. B.1-a. Payback if: guilt prevails, α is high.

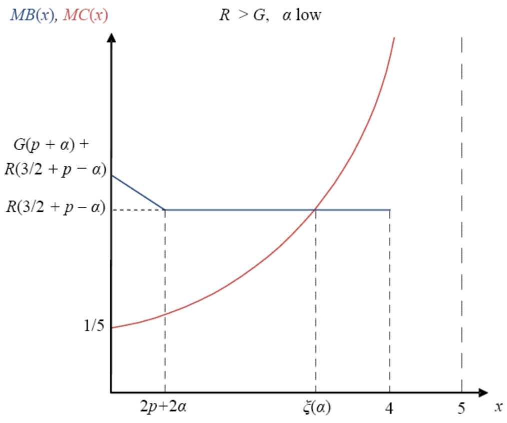

Fig. B.1-b. Payback if: recipr. prevails, α is low.

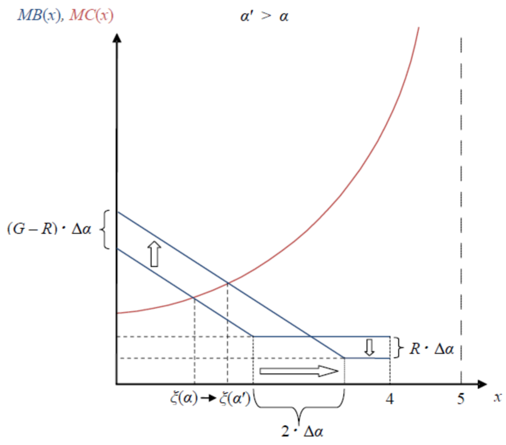

Fig. B.1-c. Payback increases if α increases.

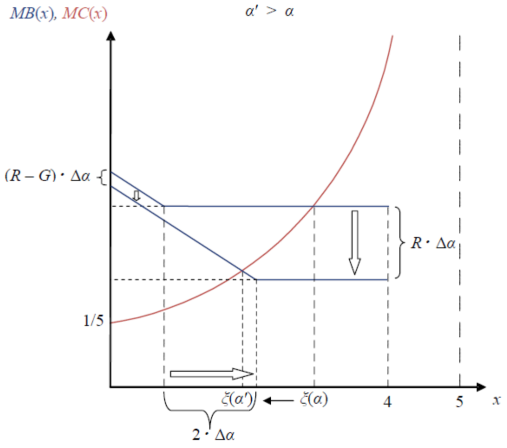

Fig. B.1-d: Payback decreases if α increases.

downward with an increase in α ($G < R$, Figure B.1-d), or the MC schedule intersects the MB schedule in its flat part (Figure B.1-b) for every α. The second condition is $R(3/2 + p - \alpha) \geq 1/(5 - 2p - 2\alpha)$ for all $\alpha \in [0, 1]$, i.e. $R \geq 1/[(3 - 2p)(1/2 + p)]$: the reciprocity component is high enough that A's disappointment does not matter to B. This explains result (iii).

For the remaining set of parameter values, the payback function is not monotone. To understand why it has to be *U-shaped*—result (ii)—, suppose that α is very small and reciprocity considerations induce B to pay back $x \geq 2p + 2\alpha$ (condition $R(3/2 + p - \alpha) \geq 1/(5 - 2p - 2\alpha)$); since there is no disappointment/guilt at x, only reciprocity matters for small changes in α; a small increase in α induces a decrease in kindness and in the payback determined by reciprocity; further increases in α bring the payback below the (increasing) $2p + 2\alpha$ threshold, and then an increase in α makes x increase if $G > R$.

More formally, Proposition B.1.1 implies that ξ is locally increasing at α (hence it is an interior solution) if and only if $G > R$ and $0 < \xi(\alpha) < 2p + 2\alpha$, which follows from the implicit function theorem: An interior solution $x^* = \xi(\alpha) \in (0, 4)$ to (3) satisfies the first-order condition (4); differentiating it, we get

$$\xi'(\alpha) = \begin{cases} -R(5 - \xi(\alpha))^2 & \text{if} \quad \xi(\alpha) \geq 2p + 2\alpha, \\ \frac{2(5-\xi(\alpha))^2}{G(5-\xi(\alpha))^2+2}(G - R) & \text{if} \quad \xi(\alpha) < 2p + 2\alpha. \end{cases}$$

B.2 Theoretical predictions for the Trust Minigame

Here we provide a rationalizability analysis of the sequential Trust Minigame with complete and incomplete information based on forward-induction reasoning. For the case of incomplete information, we use the extension to psychological games of a solution concept for **games with payoff uncertainty**, that is, games with parametrized utility functions where players have private information about the unknown utility parameters. This is not a notion of rationalizability for Bayesian games, it is an easier and more basic procedure of iterated elimination of non-best replies justified by Battigalli & Siniscalchi (2002) in their epistemic analysis of forward-induction reasoning. The epistemic justification of the solution concepts used here for psychological games can be found in Battigalli *et al.* (2020).

Rationalizability yields sharp predictions for some dominance regions of the parameter space. Outside such regions, rationalizability gives only predictions about some aspects of beliefs and the usual best-reply relation between belief and choice, but allows for every action pair. To refine our predictions, we complement our study with an equilibrium analysis. The complete-information equilibrium part is fully covered in the main text. Here we only analyze incomplete-information equilibrium.

Complete-information rationalizability

Predictions are parametrized by the *commonly known* parameter pair $(G, R) \in [0, L]^2$. We iteratively delete pairs (s_A, α) for player A, where $\alpha = \mathbb{P}_A(\mathit{Share})$, and strategies s_B for player B. We consider pairs (s_A, α) because both arguments enter the psychological utility function of B (see Battigalli *et al.* 2020). First, to ease notation, define the best-reply correspondences for A and B:

$$r_A(\alpha) = \begin{cases} \{\mathit{Cont}\} & \text{if } \alpha > \frac{1}{2} \\ \{\mathit{Cont}, \mathit{Diss}\} & \text{if } \alpha = \frac{1}{2} \\ \{\mathit{Diss}\} & \text{if } \alpha < \frac{1}{2} \end{cases}$$

$$r_B(\beta; G, R) = \begin{cases} \{\mathit{Share}\} & \text{if } WS(\beta; G, R) > 0 \\ \{\mathit{Share}, \mathit{Take}\} & \text{if } WS(\beta; G, R) = 0 \\ \{\mathit{Take}\} & \text{if } WS(\beta; G, R) < 0 \end{cases},$$

where $\beta \in [0, 1]$ denotes the conditional second-order point belief of B about α and WS is the willingness to share function

$$WS(\beta; G, R) := G\beta^2 - 2R\beta + 3R - \ell$$

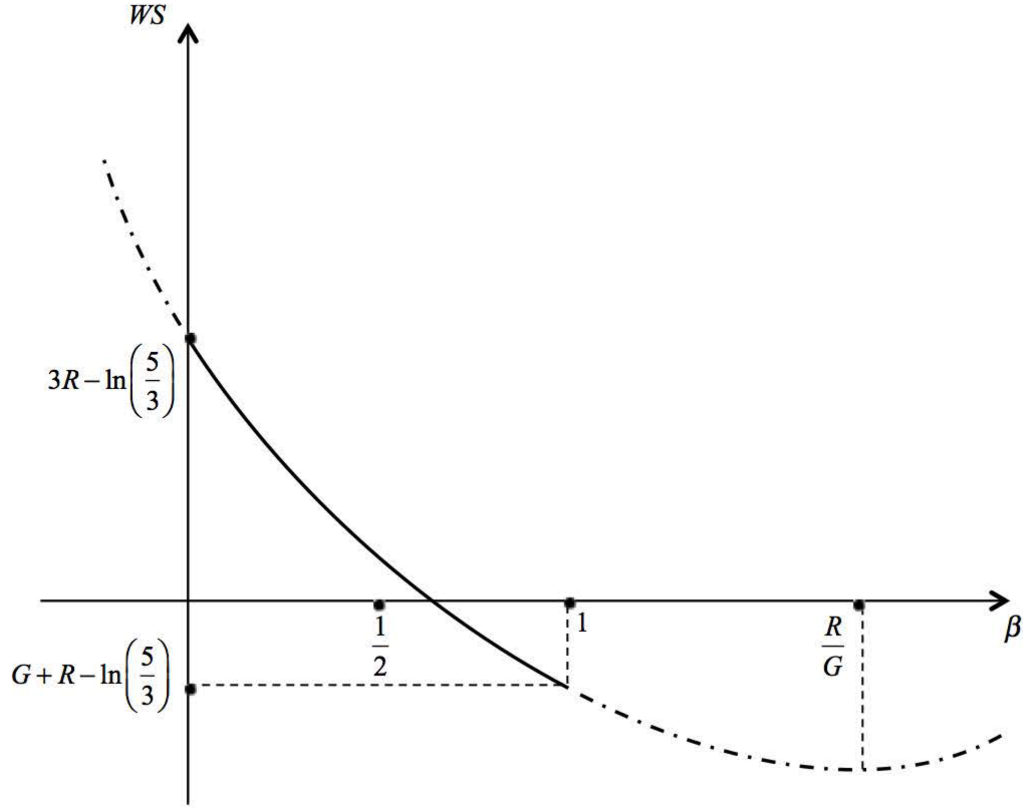

Figure B.2 B's equilibrium willingness to share for $G{+}R$ small and R/G large

obtained in eq. (5) of the paper with $\ell := \ln\left(\frac{5}{3}\right)$ (See Figure B.2). Recall that, for the purpose of this analysis, the assumption that the conditional second-order belief of B is a Dirac measure on $[0,1]$ is without loss of generality.[4]

Define step-n prediction sets $P_A^n(G,R) \subseteq S_A \times [0,1]$ and $P_B^n(G,R) \subseteq S_B$ $(n \in \mathbb{N}_0)$ recursively as follows:[5]

$$P_A^0(G,R) = S_A \times [0,1], \;\; P_B^0(G,R) = S_B,$$

$$P_A^n(G,R) = \left\{ (s_A,\alpha) : \begin{array}{l} s_A \in r_A(\alpha), P_B^{n-1}(G,R) = \{Share\} \Rightarrow \alpha = 1, \\ P_B^{n-1}(G,R) = \{Take\} \Rightarrow \alpha = 0 \end{array} \right\},$$

$$P_B^n(G,R) = \left\{ s_B \in P_B^{n-1}(G,R) : \begin{array}{l} \exists \beta \in [0,1], s_B \in r_B(\beta; G,R), \\ \left(P_{A,Cont}^{n-1}(G,R) \neq \emptyset\right) \Rightarrow \left(\beta \in P_{A,Cont}^{n-1}(G,R)\right) \end{array} \right\},$$

[4] Such result of "equivalence to certainty" holds in many decision problems. In particular, it holds for the choice between two alternatives whose utility depends continuously on an unknown variable (or parameter) that takes values in a compact and connected space.

[5] Let $P \subseteq X \times Y$ be a subset of a Cartesian product; the section at y of P is $P_y := \{x \in X : (x,y) \in P\}$. Formally, $P_B^n(G,R) = P_{B,(G,R)}$ is the section at (G,R) of the set P_B^n defined in the main text for $n \in \{1,2\}$, and later on in this appendix for any n.

where

$$P_{A,Cont}^{n-1}(G,R) := \{\alpha \in [0,1] : (Cont,\alpha) \in P_A^{n-1}(G,R)\}$$

is the section at strategy *Continue* of $P_A^{n-1}(G,R)$, and the condition on β is the forward-induction requirement.[6] In particular, $P_{A,Cont}^{1}(G,R) = [1/2,1]$; therefore, $P_B^2(G,R) = \{s_B\}$ if $r_B(\beta;G,R) = \{s_B\}$ for each $\beta \in [1/2,1]$, and $P_B^2(G,R) = S_B$ if for each $s_B \in S_B$ there is some $\beta \in [1/2,1]$ such that $s_B \in r_B(\beta;G,R)$. The complete-information rationalizable prediction set for player $i \in \{A,B\}$ is $P_i^*(G,R) = \bigcap_{n\in\mathbb{N}} P_i^n(G,R)$.

The step-1 predictions of rationalizability are given by the graphs of the best reply correspondences. The step-1 prediction set $P_A^1(G,R)$ is independent of (G,R), because A's utility does not depend on B's psychological type; $P_B^1(G,R)$ depends on whether (G,R) belongs to one of the following **simple dominance** regions of Section 3.2.1 of the paper:

$$\mathbb{S}^* := \left\{(G,R) \in [0,L]^2 : \min_{\beta\in[0,1]} WS(\beta;G,R) > 0\right\}$$

and

$$\mathbb{T}^* := \left\{(G,R) \in [0,L]^2 : \max_{\beta\in[0,1]} WS(\beta;G,R) < 0\right\}.$$

With this,

$$\begin{aligned}
P_A^1(G,R) &= \text{graph}(r_A) = \{Cont\} \times \left[\frac{1}{2},1\right] \cup \{Diss\} \times \left[0,\frac{1}{2}\right], \\
P_B^1(G,R) &= \{s_B : \exists\beta \in [0,1], s_B \in r_B(\beta;G,R)\} \\
&= \begin{cases} \{Share\} & \text{if } (G,R) \in \mathbb{S}^* \\ \{Share, Take\} & \text{if } (G,R) \notin \mathbb{S}^* \cup \mathbb{T}^* \\ \{Take\} & \text{if } (G,R) \in \mathbb{T}^* \end{cases}.
\end{aligned}$$

The last equality holds because, if (G,R) belongs to a simple dominance region, then the unique best reply is the corresponding dominant strategy independently of β, if instead (G,R) does not belong to a dominance region, then each one of the two strategies can be justified as a best reply to some conditional second-order belief β.

Before we proceed to the step-2 prediction, for the reader's convenience we recall the

[6] The forward-induction, or "best-rationalization" requirement should be

$$\left(P_{A,Cont}^{k}(G,R) \neq \emptyset\right) \Rightarrow \left(\beta \in P_{A,Cont}^{k}(G,R)\right)$$

for each $k \in \{1,...,n-1\}$, but it can be shown that this seemingly stronger condition is equivalent to the one above.

definition of the **FI-dominance** regions:

$$\mathbb{S} := \left\{ (G, R) \in [0, L]^2 : \min_{\beta \in \left[\frac{1}{2}, 1\right]} WS\left(\beta; G, R\right) > 0 \right\}$$

and

$$\mathbb{T} := \left\{ (G, R) \in [0, L]^2 : \max_{\beta \in \left[\frac{1}{2}, 1\right]} WS\left(\beta; G, R\right) < 0 \right\}.$$

Here we provide an explicit characterization of these regions. Since WS is strictly convex in β, it attains a maximum on $[1/2, 1]$ either at $\beta = 1/2$, or at $\beta = 1$. Thus, $WS(\beta; G, R) < 0$ for every $\beta \in [1/2, 1]$ if and only if $\max\{WS(1/2; G, R), WS(1; G, R)\} < 0$, where

$$\begin{aligned} WS(1/2; G, R) < 0 \quad &\Longleftrightarrow \quad \tfrac{1}{8}G + R < \tfrac{1}{2}\ell, \\ WS(1; G, R) < 0 \quad &\Longleftrightarrow \quad G + R < \ell. \end{aligned}$$

Thus (see Figure 1 in the paper),

$$\mathbb{T} = \left\{ (G, R) \in [0, L]^2 : \frac{1}{8}G + R < \frac{1}{2}\ell, G + R < \ell \right\}.$$

FI-dominance region $\mathbb{S}$ is the union of three sub-regions: Function WS attains its minimum at $\beta = R/G$, which may be to the left of $1/2$, to the right of 1, or in the interval $[1/2, 1]$; each case gives raise to a sub-region.

- If $R/G < 1/2$, WS is strictly increasing on $[1/2, 1]$, hence it attains its minimum at $\beta = 1/2$. Letting $WS(1/2; G, R) > 0$ in this case, we obtain

$$\mathbb{S}_1 := \left\{ (G, R) \in [0, L]^2 : R < \frac{1}{2}G, \frac{1}{8}G + R > \frac{1}{2}\ell \right\}.$$

- If $1/2 \leq R/G \leq 1$, WS attains its minimum in the interval $[1/2, 1]$. Letting $WS(R/G; G, R) > 0$ in this case, we obtain

$$\begin{aligned} \mathbb{S}_2 := \quad &\left\{ (G, R) \in [0, L]^2 : \tfrac{1}{2}G \leq R \leq G, WS\left(\tfrac{R}{G}; G, R\right) > 0 \right\} \\ = \quad &\left\{ (G, R) \in [0, L]^2 : \tfrac{1}{2}G \leq R \leq G, R^2 - 3GR + \ell G < 0 \right\}. \end{aligned}$$

- If $R/G > 1$, WS is strictly decreasing on $[1/2, 1]$, hence it attains its minimum at $\beta = 1$. Letting $WS(1; G, R) > 0$ in this case, we obtain $\mathbb{S}_3 := \{(G, R) \in [0, L]^2 : R > G, G + R > \ell\}$.

- Thus, $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3$ (see Figure 1 in the paper).

Then, the step-2 prediction set $P^2_{A,(G,R)}$ depends on whether (G,R) belongs to a *simple* dominance region, whereas $P^2_{B,(G,R)}$ depends on whether (G,R) belongs to an *FI*-dominance region:

$$
\begin{aligned}
P^2_A\left(G,R\right) &= \begin{cases} \left\{\left(Cont,1\right)\right\} & \text{if } (G,R) \in \mathbb{S}^* \\ P^1_A\left(G,R\right) & \text{if } (G,R) \notin \mathbb{S}^* \cup \mathbb{T}^* \\ \left\{\left(Diss,0\right)\right\} & \text{if } (G,R) \in \mathbb{S}^* \end{cases}, \\
P^2_B\left(G,R\right) &= \left\{s_B : \exists\beta\left[\frac{1}{2},1\right], s_B \in r_B\left(\beta; G,R\right)\right\} \\
&= \begin{cases} \left\{Share\right\} & \text{if } (G,R) \in \mathbb{S} \\ \left\{Share, Take\right\} & \text{if } (G,R) \notin \mathbb{S} \cup \mathbb{T} \\ \left\{Take\right\} & \text{if } (G,R) \in \mathbb{T} \end{cases}.
\end{aligned}
$$

To understand the characterization of $P^2_B\left(G,R\right)$, note that *Continue* is justified as a best reply to $\alpha \geq 1/2$, therefore $P^1_{A,Cont}\left(G,R\right) = \left[\frac{1}{2},1\right]$, which in turn implies that each $s_B \in P^n_B\left(G,R\right)$ must be a best reply to some $\beta \geq 1/2$.

Finally, the step-3 prediction sets are

$$
\begin{aligned}
P^3_A\left(G,R\right) &= \begin{cases} \left\{\left(Cont,1\right)\right\} & \text{if } (G,R) \in \mathbb{S} \\ P^2_A\left(G,R\right) & \text{if } (G,R) \notin \mathbb{S} \cup \mathbb{T} \\ \left\{\left(Diss,0\right)\right\} & \text{if } (G,R) \in \mathbb{S} \end{cases} \\
&= \begin{cases} \left\{\left(Cont,1\right)\right\} & \text{if } (G,R) \in \mathbb{S} \\ P^1_A\left(G,R\right) & \text{if } (G,R) \notin \mathbb{S} \cup \mathbb{T} \\ \left\{\left(Diss,0\right)\right\} & \text{if } (G,R) \in \mathbb{S} \end{cases}, \\
P^3_B\left(G,R\right) &= P^2_B\left(G,R\right).
\end{aligned}
$$

The second equality for $P^3_A\left(G,R\right)$ follows from the characterization of $P^2_A\left(G,R\right)$ and the fact that $\mathbb{S}^* \subseteq \mathbb{S}$ and $\mathbb{T}^* \subseteq \mathbb{T}$. The equality for $P^3_B\left(G,R\right)$ holds because step 3 may yield further restrictions on conditional second-order beliefs,[7] but this does not imply further restrictions on behavior.[8] This implies that $P^n_i\left(G,R\right) = P^3_i\left(G,R\right)$ for each i and $n > 3$. Hence, $P^*_i\left(G,R\right) = P^3_i\left(G,R\right)$ is the complete-information rationalizable prediction for i given common knowledge of (G,R). To summarize, we recall Proposition 1 of the paper:

Proposition B.2.1 *Under complete information, the prediction of rationalizability based on forward induction is as follows:*

[7] If $(G,R) \in \mathbb{S}$ then $\beta = 1$, because B is certain—both initially and conditionally on *Continue*—that A expects him to share.

[8] If $(G,R) \in \mathbb{S}$, then $s_B = Share$, as in the step 2.

(i) Continue*,* Share*, and* $\alpha = 1$ *if* $(G, R) \in \mathbb{S}$*,*

(ii) Dissolve*,* Take*, and* $\alpha = 0$ *if* $(G, R) \in \mathbb{T}$*,*

(iii) any (s_A, s_B, α) *such that* s_A *is a best reply to* α *(i.e.,* $(s_A, \alpha) \in P_A^1$*) is possible if* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$*.*

Incomplete-information rationalizability

Under incomplete information, A does not know B's type (G, R); therefore, only the predictions for B depend on (G, R). To obtain a prediction about s_B player A forms beliefs about the relationship between s_B and (G, R); hence, A holds a belief μ about B's strategy and psychological type and takes a best reply to her marginal belief about s_B, viz. $(\alpha, 1 - \alpha) = \text{marg}_{S_B}\mu$, where α is the probability of *Share*. But—absent restrictions on A's exogenous belief about (G, R)—this does not allow to say much about A. Taking A's perspective, the prediction set concerning B is a subset of $S_B \times [0, L]^2$, where L is the commonly known upper bound on G and R. The step-n prediction sets are recursively defined as follows:

$$P_A^0 = S_A \times [0, 1], \;\; P_B^0 = S_B \times [0, L]^2,$$

$$P_A^n = \left\{(s_A, \alpha) : s_A \in r_A(\alpha), \exists \mu \in \Delta\left(P_B^{n-1}\right), (\alpha, 1-\alpha) = \text{marg}_{S_B}\mu\right\},$$

$$P_B^n = \left\{(s_B; G, R) \in P_B^{n-1} : \exists \beta \in [0, 1], s_B \in r_B\left(\beta; G, R\right), \left(P_{A,Cont}^{n-1} \neq \emptyset\right) \Rightarrow \left(\beta \in P_{A,Cont}^{n-1}\right)\right\}.$$

The rationalizable prediction set for player i is $P_i^* = \bigcap_{n \geq 1} P_i^n$. With this, it is easy to derive these prediction sets and relate them with the complete-information sets:

- $P_A^n = \{(s_A, \alpha) : s_A \in r_A(\alpha)\}$ for each $n \geq 1$, thus

$$P_A^* = \{Cont\} \times \left[\frac{1}{2}, 1\right] \cup \{Diss\} \times \left[0, \frac{1}{2}\right];$$

- for $n = 1, 2$ the complete-information prediction $P_{B,Cont}^n(G, R)$ is the section at (G, R) of the incomplete-information prediction P_B^n, thus

$$\begin{aligned} P_B^1 &= \{Share\} \times \mathbb{S}^* \cup \{Take\} \times \mathbb{T}^* \cup \{Share, Take\} \times \left([0, L]^2 \backslash \left(\mathbb{S}^* \cup \mathbb{T}^*\right)\right), \\ P_B^2 &= \{Share\} \times \mathbb{S} \cup \{Take\} \times \mathbb{T} \cup \{Share, Take\} \times \left([0, L]^2 \backslash \left(\mathbb{S} \cup \mathbb{T}\right)\right); \end{aligned}$$

- $P_B^n = P_B^2$ for every $n > 2$, thus

$$P_B^* = \{Share\} \times \mathbb{S} \cup \{Take\} \times \mathbb{T} \cup \{Share, Take\} \times \left([0, L]^2 \backslash \left(\mathbb{S} \cup \mathbb{T}\right)\right).$$

In words, absent any hypothesis about A's exogenous beliefs, incomplete-information rationalizability only predicts that A chooses a best reply to whatever first-order belief α she holds, and that B chooses according to the FI-dominance regions. To summarize, we recall Proposition 3 of the paper:

Proposition B.2.2 *Without restrictions on exogenous beliefs, incomplete-information rationalizability implies (only) that* $(s_A, \alpha) \in P_A^1$ *and* $(s_B; G, R) \in P_B^2$*; in particular,* B *chooses* Share *if* $(G, R) \in \mathbb{S}$ *and* Take *if* $(G, R) \in \mathbb{T}$*, while both strategies are rationalizable for* $(G, R) \notin \mathbb{S} \cup \mathbb{T}$.

Incomplete-information equilibrium

To give more structure to the incomplete-information predictions and for comparability with the complete-information analysis, we introduce assumptions about players' exogenous beliefs following Harsanyi's methodology and then we perform a Bayesian-equilibrium analysis (cf. Attanasi *et al.* 2016).

To obtain a (psychological) Bayesian game, we append an exogenous type structure to the game with payoff uncertainty (cf. Harsanyi 1967-68). This yields an implicit specification of the possible **hierarchies of exogenous beliefs** of A and B about the psychological type $(G, R) \in [0, L]^2$: a first-order belief of A over $[0, L]^2$, a second-order belief of B about such first-order belief, and so on. Formally, a type structure is given by compact, metrizable **type spaces** $\mathcal{T}_i$ and continuous[9] **belief maps** $\tau_i : \mathcal{T}_i \to \Delta(\mathcal{T}_{-i})$ $(i = A, B)$, where $\mathcal{T}_B = [0, L]^2 \times \mathcal{E}_B$. Thus, a **type** of B is a triple $t_B = (G, R, e_B)$ where (G, R) is the psychological type and e_B is the **epistemic type**. The latter determines the beliefs of B about the beliefs of A. Since the utility function of A is commonly known, A's type is purely epistemic, and it determines the beliefs of A about (G, R) and the beliefs of B. The (exogenous) first-order belief of type t_A about (G, R) is $\tau_A^1(t_A) = \text{marg}_{[0,L]^2} \tau_A(t_A)$. The belief of t_B about $\tau_A^1(t_A)$ is the exogenous second-order belief of t_B, $\tau_B^2(t_B)$, and so on. We postpone the details about the type structure.

A **Bayesian equilibrium** is a pair of measurable decision functions $\sigma_A : \mathcal{T}_A \to \{Cont, Diss\}$ and $\sigma_B : \mathcal{T}_B \to \{Share, Take\}$ such that, for all i and t_i, $\sigma_i(t_i)$ is at least as good as the alternative, given the beliefs of t_i about the action and beliefs of the co-player. Such **endogenous** beliefs are derived from the exogenous belief maps and the decision functions, as we show below. For now, it suffices to say that each type t_A holds a first-order belief

[9] $\Delta(\mathcal{T}_{-i})$ is endowed with the topology of weak convergence of Borel probability measures: $\mu_i^n \to \mu_i$ if and only if $\int_{\mathcal{T}_{-i}} f(t_{-i}) \mu_i^n(\mathrm{d}t_{-i}) \to \int_{\mathcal{T}_{-i}} f(t_{-i}) \mu_i(\mathrm{d}t_{-i})$ for every continuous (hence bounded) function $f : \mathcal{T}_{-i} \to \mathbb{R}$.

$\alpha(t_A) = \mathbb{P}_{t_A}(Share)$, and $\sigma_A(t_A) = Cont$ only if $\alpha(t_A) \geq 1/2$. Each type t_B holds a second-order cdf $\mathbb{P}_{t_B}\left(\widetilde{\alpha} \leq x | Cont\right)$ provided that $\tau_B(t_B)\left(\{t_A : \sigma_A(t_A) = Cont\}\right) > 0$, which implies $\tau_B(t_B)\left(\{t_A : \alpha(t_A) \geq 1/2\}\right) > 0$.

We assume that a set with (strictly) positive measure of A's epistemic types assigns more than 50% probability to the simple dominance region $\mathbb{S}^*$. Given a full support restriction on the exogenous beliefs of B, this implies that in equilibrium B predicts that A chooses *Continue* with strictly positive probability and has well-defined conditional beliefs given *Continue*; with this, the forward-induction analysis developed above applies to such Bayesian equilibria (which are also perfect).

Next we provide the details of the type structure $(\mathcal{T}_i, \tau_i)_{i\in\{A,B\}}$:

1. $\mathcal{T}_A = \times_{i=1}^{i=d_A}[0, \bar{t}_A^i] \subseteq \mathbb{R}^{d_A}$.

2. $\mathcal{T}_B = [0, L]^2 \times \mathcal{E}_B$, where $\mathcal{E}_B = \times_{j=1}^{j=d_B}[0, \bar{e}_A^j] \subseteq \mathbb{R}^{d_B}$ (the dimensions $d_A, d_B \in \mathbb{N}$ may be different).

3. $\tau_A : \mathcal{T}_A \to \Delta\left(\mathcal{T}_B\right)$ is *continuous* and such that, for every $(G, R) \in [0, L]^2$ and open set $\emptyset \neq O_B \subseteq \mathcal{E}_B$,

$$\tau_A(t_A)\left([0, G] \times [0, R] \times O_B\right) = F_{t_A}(G, R)\mu(O_B) > 0,$$

   where $\mu \in \Delta(\mathcal{E}_B)$ is absolutely continuous with respect to the Lebesgue measure, with *density* function *bounded away from zero.* Furthermore,

$$\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\} \neq \emptyset.$$

4. $\tau_B : \mathcal{T}_A \to \Delta\left(\mathcal{T}_B\right)$ is *continuous* and such that $\tau_B(G, R, e_B)$ depends only on e_B (hence we write $\tau_B(e_B)$ to ease notation); furthermore, for every $e_B \in \mathcal{E}_B$, $\tau_B(e_B)$ is absolutely continuous with respect to the Lebesgue measure, with *density* function *bounded away from zero.*

The last assumption implies that, for every open set $\emptyset \neq O_A \subseteq \mathcal{T}_A$ and every $e_B \in \mathcal{E}_B$, $\tau_B(e_B)(O_A) > 0$. Since the belief map τ_A is continuous and each $\tau_B(e_B)$ is absolutely continuous, the set of types

$$\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\}$$

is open.[10] By assumption, this set is also nonempty, therefore

$$\tau_B(e_B)\left(\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\}\right) > 0$$

for each $e_B \in \mathcal{E}_B$. (A similar argument shows that $\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{T}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\}$ is open, but we do not assume that it is nonempty.)

Next we describe the qualitative features of every Bayesian equilibrium of the model.[11] First note that if $t_B = (G, R, e_B)$ is such that $(G, R) \in \mathbb{S}^*$, then necessarily $\sigma\left(t_B\right) = Share$. Therefore, for each $t_A \in \left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\}$,[12]

$$\alpha\left(t_A\right) := \tau_A\left(t_A\right)\left(\sigma_B^{-1}\left(Share\right)\right) \geq \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2},$$

which implies $\sigma_A\left(t_A\right) = Cont$, that is,

$$\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\} \subseteq \sigma_A^{-1}\left(Cont\right).$$

Assumptions 3 and 4 imply that, for each $e_B \in \mathcal{E}_B$,

$$\begin{aligned}\tau_B\left(e_B\right)\left(\sigma_A^{-1}\left(Cont\right)\right) &= \tau_B\left(e_B\right)\left(\left\{t_A \in \mathcal{T}_A : \alpha(t_A) > \frac{1}{2}\right\}\right) \\ &\geq \tau_B(e_B)\left(\left\{t_A \in \mathcal{T}_A : \tau_A(t_A)\left(\mathbb{S}^* \times \mathcal{E}_B\right) > \frac{1}{2}\right\}\right) > 0.\end{aligned}$$

Hence, the equilibrium conditional belief

$$\tau_B(e_B)(\cdot|Cont) := \tau_B(e_B)\left(\cdot|\sigma_A^{-1}\left(Cont\right)\right)$$

is well defined for each e_B, and

$$\tau_B(e_B)\left(\widetilde{\alpha} \geq \frac{1}{2}\middle| Cont\right) = \tau_B(e_B)\left(\left\{t_A \in \mathcal{T}_A : \alpha(t_A) \geq \frac{1}{2}\right\}|\sigma_A^{-1}\left(Cont\right)\right) = 1.$$

[10] The boundary of $\mathbb{S} \times \mathcal{E}_B$ has zero probability for each type t_A. Hence, the portmanteau theorem and continuity of τ_A with respect to the weak convergence of measures imply that, for each converging sequence $t_A^n \to \bar{t}_A$, $\lim_{n\to\infty} \tau_A(t_A^n)\left(\mathbb{S} \times \mathcal{E}_B\right) = \tau_A(\bar{t}_A)\left(\mathbb{S} \times \mathcal{E}_B\right)$. This in turn implies that, for each $x \in [0, 1]$, the set of types t_A such that $\tau_A(t_A)\left(\mathbb{S} \times \mathcal{E}_B\right) > x$ is open.

[11] For an incomplete-information model without reciprocity, we can prove the existence and uniqueness of such equilibrium. We cannot do the same here without adding more structure.

[12] We use standard notation for the set of preimages of the value of a function. In particular, $\sigma_i^{-1}\left(s_i\right) = \left\{t_i \in \mathcal{T}_i : \sigma_i\left(t_i\right) = s_i\right\}$.

Therefore, for every B-type (G, R, e_B),

$$\begin{aligned} (G,R) \in \mathbb{S} \quad &\Rightarrow \quad \sigma_B(G,R,e_B) = Share, \\ (G,R) \in \mathbb{T} \quad &\Rightarrow \quad \sigma_B(G,R,e_B) = Take. \end{aligned}$$

Since each type space is connected and each belief map is continuous, it follows that every value between the minimum and maximum of $\alpha(t_A)$ is attained by some t_A, and every value between the minimum and maximum of $\beta^0(e_B) := \mathbb{E}_{e_B}[\widetilde{\alpha}]$ is attained by some e_B:

$$\begin{aligned} \alpha\left(\mathcal{T}_A\right) &= \left[\min_{t_A \in \mathcal{T}_A} \alpha(t_A), \max_{t_A \in \mathcal{T}_A} \alpha(t_A)\right], \\ \beta^0(\mathcal{E}_B) &= \left[\min_{e_B \in \mathcal{E}_B} \beta^0(e_B), \max_{e_B \in \mathcal{E}_B} \beta^0(e_B)\right]. \end{aligned}$$

The belief of every type t_A—including the minimizers of $\tau_A(t_A)\,(\mathbb{S}\times\mathcal{E}_B)$ and $\tau_A(t_A)\,(\mathbb{T}\times\mathcal{E}_B)$—is determined by a probability density function bounded away from zero; hence,

$$\begin{aligned} \min_{t_A \in \mathcal{T}_A} \alpha(t_A) &\geq \min_{t_A \in \mathcal{T}_A} \tau_A(t_A)\,(\mathbb{S}\times\mathcal{E}_B) > 0, \\ \max_{t_A \in \mathcal{T}_A} \alpha(t_A) &\leq 1 - \min_{t_A \in \mathcal{T}_A} \tau_A(t_A)\,(\mathbb{T}\times\mathcal{E}_B) < 1. \end{aligned}$$

This in turn implies:

$$\begin{aligned} \min_{e_B \in \mathcal{E}_B} \beta^0(e_B) &\geq \min_{t_A \in \mathcal{T}_A} \alpha(t_A) > 0, \\ \max_{e_B \in \mathcal{E}_B} \beta^0(e_B) &\leq \max_{t_A \in \mathcal{T}_A} \alpha(t_A) < 1. \end{aligned}$$

The following proposition summarizes these observations.

Proposition B.2.3 *Every Bayesian equilibrium of the model has the following features:*

(a) For every $t_A \in \mathcal{T}_A$,

$$\begin{aligned} \tau_A(t_A)\,(\mathbb{S} \times \mathcal{E}_B) > \tfrac{1}{2} \quad &\Rightarrow \quad \sigma_A(t_A) = Cont, \\ \tau_A(t_A)\,(\mathbb{T} \times \mathcal{E}_B) > \tfrac{1}{2} \quad &\Rightarrow \quad \sigma_A(t_A) = Diss, \end{aligned}$$

and

$$\tau_A(t_A)\,(\mathbb{S} \times \mathcal{E}_B) \leq \alpha(t_A) \leq 1 - \tau_A(t_A)\,(\mathbb{T} \times \mathcal{E}_B)\,.$$

(b) For every $(G,R,e_B) \in \mathcal{T}_B$,

$$\begin{aligned}(G,R) \in \mathbb{S} \quad &\Rightarrow \quad \sigma_B(G,R,e_B) = \mathit{Share},\\ (G,R) \in \mathbb{T} \quad &\Rightarrow \quad \sigma_B(G,R,e_B) = \mathit{Take},\end{aligned}$$

and

$$\tau_B(e_B)\left(\widetilde{\alpha} \geq \frac{1}{2}\middle| \mathit{Cont}\right) = 1,$$

hence $\beta(e_B) \geq 1/2$.

(c) There are values $\underline{\alpha} > 0$ *and* $\bar{\alpha} < 1$ *such that,*

$$\underline{\alpha} \leq \alpha(t_A), \beta^0(e_B) \leq \bar{\alpha},$$

for every $t_A \in \mathcal{T}_B$ *and* $e_B \in \mathcal{E}_B$.

(d) All values between $\min_{t_A \in \mathcal{T}_A} \alpha(t_A)$ *and* $\max_{t_A \in \mathcal{T}_A} \alpha(t_A)$ *are attained by some* t_A*, and all values between* $\min_{e_B \in \mathcal{E}_B} \beta^0(e_B)$ *and* $\max_{e_B \in \mathcal{E}_B} \beta^0(e_B)$ *are attained by some* e_B.

Of course, we need to add assumptions about the actual distribution of types in order to derive from the equilibrium analysis implications about the distribution of behavior and endogenous beliefs. We assume that the distribution of types is such that the type of A and the type of B are statistically independent, the epistemic type of B is independent of the psychological type of B,[13] and the marginal distributions are determined by density functions bounded away from zero: For all $\breve{t}_A \in \mathcal{T}_A$, $(\breve{G},\breve{R}) \in [0,L]^2$, $\breve{e}_B \in \mathcal{E}_B$,

$$\begin{aligned}&\mathbb{P}\left([0,\breve{t}_A] \times [0,\breve{G}] \times [0,\breve{R}] \times [0,\breve{e}_B]\right) =\\ &= \left(\int_{[0,\breve{t}_A]} f_A(t_A)\mathrm{d}t_A\right) \times \left(\int_{[0,\breve{G}]\times[0,\breve{R}]} f_B^1(G,R)\mathrm{d}G\mathrm{d}R\right) \times \left(\int_{[0,\breve{e}_B]} f_B^2(e_B)\mathrm{d}e_B\right),\end{aligned}$$

where the densities f_A, f_B^1 and f_B^2 are bounded away from zero, respectively, on $\mathcal{T}_A$, $[0,L]^2$ and $\mathcal{E}_B$.[14] From this, one can derive the qualitative predictions about behavior and endogenous beliefs of Proposition 4 of the paper. In particular, the positive correlation between the fraction of B-types with $G \geq 2R$ choosing *Share* and the conditional second-order beliefs holds if the psychological type and epistemic type of B are statistically independent. This, however, is just a sufficient condition to obtain such positive correlation.

[13] Assumption 3 about the type structure contains the corresponding independence condition for the beliefs of A about B.

[14] Recall that, in the present model, t_A and e_B are vectors in $\mathbb{R}^{d_A}$ and $\mathbb{R}^{d_B}$, respectively. Hence $[0,\breve{t}_A]$ and $[0,\breve{e}_B]$ are Cartesian products of intervals, and f_A, f_B^2 are multivariate density functions.