Biased-Belief Equilibrium: Online Appendix

Yuval Heller* Eyal Winter†

February 21, 2019

Abstract

These online appendices supplement the paper "biased-belief equilibrium" published in the *American Economic Journal: Microeconomics*. Appendix A presents various interesting examples. We formally present the evolutionary interpretation of our solution concept in Appendix B, and the delegation interpretation in Appendix C. Appendix D relaxes the assumption that biased beliefs have to be continuous. Appendix E shows how to extend our results to a setup with partial observability. Appendix F presents our formal proofs.

A Additional Examples

A.1 A Non-Nash Strong BBE Outcome in a Zero-Sum Game

The following example shows that although the weak BBE payoff must be the Nash equilibrium payoff in a zero-sum game, the strategy profile sustaining it need not be a Nash equilibrium.

Example 6. Consider the symmetric rock–paper–scissors zero-sum game described in Table 2. We show that $\left(\left(I_d, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right), \left(R, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right)\right)$ is a strong BBE, in which the player 1 (he) has

Table 2: Symmetric Rock-Paper-Scissors Zero-Sum Game Payoffs

	R	P	S
R	0, 0	0,1	1,0
P	1,0	0, 0	0,1
S	0,1	1,0	0, 0

undistorted beliefs and plays R, while player 2 (she) has a blind belief that the opponent always mixes equally, and she mixes equally. It is immediate that $\left(R, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right) \in NE\left(G_{\left(I_d, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right)}\right)$, and the equilibrium payoff to each player is zero. Next, observe that after any deviation of player 1 to a biased belief ψ_1', there is an equilibrium of the game $G_{\left(\psi_1', \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right)}$ in which player 2 mixes equally and player 1 obtains a payoff of zero. Finally, observe that after any deviation of player 2 to a biased belief ψ_1', player 1 obtains a payoff of at least zero (her minmax payoff in $G_{\left(I_d, \psi_2'\right)}$) in any Nash equilibrium in $G_{\left(I_d, \psi_2'\right)}$, which implies that player 2 obtains a payoff of at most zero, and, as a result, she does not gain from the deviation.

*Department of Economics, Bar Ilan University, Israel. yuval.heller@biu.ac.il. URL: https://sites.google.com/site/yuval26/.

†Lancaster University, Management School, UK, and the Hebrew University, Department of Economics, Israel. mseyal@mscc.huji.ac.il . URL: http://www.ma.huji.ac.il/∼mseyal/.

A.2 Prisoner's Dilemma with a Weakly Dominated Withdrawal Strategy

Proposition 2 implies, in particular, that defection is the unique weak BBE outcome in the prisoner's dilemma game. The following example demonstrates that a relatively small change to the prisoner's dilemma game, namely, adding a third weakly dominated "withdrawal" strategy that transforms "cooperation" into a weakly dominated strategy, can allow us to sustain cooperation as a strong BBE outcome. This is done by means of biases under which a player believes that his opponent is planning to withdraw from the game whenever he intends to cooperate, which makes cooperation a rational move.

Example 7. Consider the variant of the prisoner's dilemma game with a third "withdrawal" action as described in Table 3. In this symmetric game both players get a high payoff of 10 if they both

Table 3: Prisoner's Dilemma Game with a Withdrawal Action

	c	d	w
c	10,10	0,11	0,0
d	11,0	1,1	0,0
w	0,0	0,0	0,0

play action c (interpreted as cooperation). If one player plays d (*defection*) and his opponent plays c, then the defector gets 11 and the cooperator gets 0. If both players defect, then each of them gets a payoff of 1. Finally, if either player plays action w (interpreted as *withdrawal*), then both players get 0. Observe that defection is a weakly dominant action, and that the game admits two Nash equilibria: (w, w) and (d, d), inducing respective symmetric payoffs of zero and one.

We identify a mixed action with a vector $(\alpha_c, \alpha_d, \alpha_w)$, where $\alpha_c \geq 0$ (resp., $\alpha_d \geq 0$, $\alpha_w \geq 0$) denotes the probability of choosing action c (resp., d, w). For each player i, let ψ_i be the following biased-belief function:

$$\psi_i^* (\alpha_c, \alpha_d, \alpha_w) = (0, \alpha_d, \alpha_c + \alpha_w) .$$

We now show that $((\psi_1^*, \psi_2^*), (c, c))$ is a non-monotone strong BBE in which both players obtain a high payoff of 10 (which is strictly better than the best Nash equilibrium payoff, and strictly better than the Stackelberg payoff of each player). Observe first that $c \in BR(\psi_i^*(c)) = BR(w)$, which implies that $(c, c) \in NE\left(G_{(\psi_1^*, \psi_2^*)}\right)$. Next, consider a deviation of player i to biased belief ψ_i'. Observe that player i can gain a payoff higher than 10 only if he plays action d with positive probability, but this implies that the unique best reply of player j to his biased belief about player i's strategy is defection, which implies that player i obtains a payoff of at most one.

A.3 The Folk Theorem Result Does not Hold for All Finite Games

The following example demonstrates that the folk theorem result (Proposition 4) does not necessarily hold for games that do not admit best replies with full undominated support.

Example 8. Consider the three-action symmetric game described in Table 4. Observe that all the actions in the game are undominated, and that the game does not admit best replies with full undominated support: there is no strategy of the opponent for which one of the players has a best

Table 4: A Game in which (a,a) is not a Monotone Weak BBE Outcome

	a	b	c
a	2, 2	2, 3	1.1, 3
b	3, 2	3, 3	1, 0
c	3.1, 1	0, 1	0, 0

reply with full support. This is so because action a (c) is a best reply only to his opponent's strategies that assign a probability of at least 90% to action c (a), which implies that actions a and c cannot be best replies simultaneously. Observe that the undominated minmax payoff of each player is equal to 1 (because the opponent can play the undominated action c, and by playing this the opponent guarantees that the player gets a payoff of at most 1).

Consider the undominated action profile (a,a) (which induces a payoff strictly above the undominated minmax payoff to each player). We will show that (a,a) is not a monotone weak BBE (which demonstrates that the folk theorem result of Proposition 8 does not hold in this game). Assume to the contrary that (a,a) is a monotone weak BBE. Let $\left(\left(\psi_1^*,\psi_2^*\right),(a,a)\right)$ be a monotone weak BBE. The fact that $(a,a)\in NE\left(G_{\left(\psi_1^*,\psi_2^*\right)}\right)$ implies that $\psi_1^*(a)(c)>90\%$. Consider a deviation of player 2 to having the blind belief $\psi_2'=b$. Observe that player 2 plays action b in any equilibrium of $G_{\left(\psi_1^*,\psi_2'\right)}$. The monotonicity of ψ_1^* implies that $\psi_1^*(b)(a)\leq\psi_1^*(a)(a)\leq 1-\psi_1^*(a)(c)\leq 10\%$, which implies that the best reply of player 1 to the perceived strategy of player 2 ($\psi_1^*(b)$) does not have action c in its support. This implies that player 1 gains a payoff of at least 3 in any Nash equilibrium of the new biased game $G_{\left(\psi_1^*,\psi_2'\right)}$, which contradicts $\left(\left(\psi_1^*,\psi_2^*\right),(a,a)\right)$ being a monotone weak BBE.

A.4 Examples of Games with Strategic Complements

In this subsection we analyze three examples of games with strategic complements: input games, stag hunt games, and the traveler's dilemma.

Our first example demonstrates how to implement the undominated Pareto optimal profile as a strong BBE in an input (or partnership game).

Example 9 (*Input games*)**.** Consider the following input game (closely related games are analyzed in, among others, Holmstrom, 1982 and Heller and Sturrock, 2017). Let $S_i=S_j=[0,1]$, and let the payoff function be $\pi_i(s_i,s_j,\rho)=s_i\cdot s_j-\frac{s_i^2}{2\rho}$, where the parameter $\frac{1}{\rho}$ is interpreted as the cost of effort. One can show that (1) the best-reply function of each agent is to exert an effort that is $\rho<1$ times smaller than the opponent's (i.e., $BR(s_j)=\rho\cdot s_j$), (2) in the unique Nash equilibrium each player exerts no effort $s_i=s_j=0$, (3) the highest undominated strategy of each player i is $s_i=\rho$, and (4) the undominated strategy profile (ρ,ρ) is Nash improving and yields the best payoff to both players out of all the undominated symmetric strategy profiles. Let ψ_i^* be the following biased-belief function:

$$\psi_i^*(s_j)=\begin{cases}\frac{s_j}{\rho} & s_j<\rho\\ 1 & s_j\geq\rho.\end{cases}$$

Observe that ψ_i^* is monotone and exhibits wishful thinking. We now show that $\left(\left(\psi_1^*,\psi_2^*\right),(\rho,\rho)\right)$ is a strong BBE. Observe that $BR(\psi_i^*(s_j))=BR\left(\frac{s_j}{\rho}\right)=s_j$ for any $s_j\leq\rho$, and that $BR(\psi_i^*(s_j))=$

$BR(1) = \rho$ for any $s_j \geq \rho$. This implies that $(\rho, \rho) \in NE\left(G_{(\psi_1^*, \psi_2^*)}\right)$, and that for any player i, any biased belief ψ_i', and any Nash equilibrium (s_1', s_2') of the biased game $G_{(\psi_i', \psi_j)}$, $s_j' = min(s_i', \rho)$. This implies that $\pi_i(s_1', s_2') \leq \pi_i(\rho, \rho)$, which shows that $((\psi_1^*, \psi_2^*), (\rho, \rho))$ is a strong BBE. Observe that this BBE induces only a small distortion in the belief of each player, assuming that ρ is sufficiently close to one:

$$|\psi_i^*(s_j) - s_j| < \left|\frac{s_j}{\rho} - s_j\right| < \frac{1-\rho}{\rho}.$$

Our second example characterizes the set of BBE outcomes (and their supporting beliefs) in stag hunt games.

Example 10 (*Stag hunt games*)**.** Stag hunt is a two-action game describing a conflict between safety and social cooperation. Specifically, each player i has two actions: s_i ("stag") and h_i ("hare"), and his ordinal preferences are $(s_i, s_j) \succ_i (h_i, s_j) \succeq_i (h_i, h_j) \succ_i (s_i, h_j)$. Table 5 presents the payoff of a

Table 5: Stag Hunt Game ($g_1, g_2 \in (0, 1]$ and $l_1, l_2 > 0$)

	s_2	h_2
s_1	$1, 1$	$-l_1, g_2$
h_1	$g_1, -l_1$	$0, 0$

general stag hunt game, where we have normalized, without loss of generality, the payoff of each player when playing action profile (s_i, s_j) $((h_i, h_j))$ to be one (zero), and where each g_i is positive and each l_i is in the interval $(0, 1)$. A common interpretation of stag hunt games (à la Jean-Jacques Rousseau) is a situation in which two individuals go hunting. Each can individually choose to hunt a stag or to hunt a hare. Each player must choose an action without knowing the choice of the other. If an individual hunts a stag, he must have the cooperation of his opponent in order to succeed. An individual can get a hare by himself, but a hare is worth less than a stag. It is well known that the game admits 3 equilibria: (s_i, s_j), (h_i, h_j), and (α_1^*, α_2^*), with

$$\alpha_i^* = \frac{l_j}{l_j + (1 - g_j)} \in (0, 1),$$

where each α_i represents the probability that player i plays s_i.

Applying the analysis of the previous section shows that the game admits 3 classes of BBE:

- Hunting the hare:$((\psi_1^*, \psi_2^*), (0, 0))$, where each ψ_i^* is an arbitrary monotone biased belief that satisfies $\psi_i^*(1) \geq \alpha_i^*$.
- Hunting the stag. $((\psi_1^*, \psi_2^*), (1, 1))$, where each ψ_i^* is an arbitrary monotone biased belief that satisfies $\psi_i^*(1) \leq \alpha_i^*$.
- Mixing with less weight to hunting the stag, wishful thinking, and responsiveness to bad news: $((\psi_1^*, \psi_2^*), (\beta_1, \beta_2))$, where for each player i: (1) the payoff is above the minmax payoff: $\pi_i(\beta_i, \beta_j) \geq 0$, (2) the players hunt the stag less often in the unique Nash equilibrium: $\beta_i \in$

$(0, \alpha_i^*)$, (3) wishful thinking: $\psi_i^*(\beta_j) = \alpha_j^* > \beta_j$, (4) responsiveness to bad news: $\psi_i^*(\alpha) = \alpha_j^*$ for each $\alpha \geq \beta_j$, and $\psi_i^*(\alpha) < \alpha_j^*$ for each $\alpha < \beta_j$.

Observe that any profile (β_1, β_2), where $\beta_i \in (\alpha_i^*, 1)$, cannot be a BBE outcome. If $\beta_j = 1$, then player i can gain by deviating to $\psi_i' \equiv 1$, as the unique equilibrium of the new biased game is $(1, 1)$, which induces a higher payoff to player i relative to (β_i, β_j). If $\beta_j < 1$, then player j can gain by deviating to $\psi_j' \equiv 1$, as the only possible equilibria of the new biased game are $(1, 1)$ and $(\beta_i, 1)$, both of which induce a higher payoff to player j relative to (β_i, β_j).

Our third example deals with the traveler's dilemma game, in which each agent has 100 pure ordered actions that have a discrete payoff structure that resembles strategic complementarity in interval games. We demonstrate how to implement the undominated Pareto optimal profile in this game as a strong BBE outcome that presents wishful thinking.

Example 11 (*Implementing the undominated Pareto optimal profile as a strong BBE in the traveler's dilemma*)**.**

Consider the following version of the traveler's dilemma game (Basu, 1994). Each player has 100 actions ($A_i = \{1, ..., 100\}$), and the payoff function of each player is

$$\pi_i(a_i, a_j) = \begin{cases} a_i + 2 & a_i < a_j \\ a_i & a_i = a_j \\ a_j - 2 & a_i > a_j. \end{cases}$$

The interpretation of the game is as follows. Two identical suitcases have been lost, each owned by one of the players. Each player has to evaluate the value of his own suitcase. Both players get a payoff equal to the minimal evaluation (as the suitcases are known to have identical values), and, in addition, if the evaluations differ, then the player who gave the lower (higher) evaluation gets a bonus (malus) of 2 to his payoff.

It is well known that the unique Nash equilibrium is $(1, 1)$, which yields a low payoff of one to each player. Observe that the traveler's dilemma has positive spillovers, in the sense that it is always weakly better for a player if his opponent chooses a higher action. The traveler's dilemma has strategic complementarity in the sense that the best reply of an agent is to stop one stage before his opponent, and, thus, an agent has an incentive to choose a higher action if his opponent chooses a higher action.

Observe that action 99 is the "highest" undominated action of each player (as 99 is a best reply against 100, and as action 100 is not a best reply against any of the opponent's strategies). In what follows, we construct a strong BBE exhibiting wishful thinking that yields a payoff of 99 to each player in the undominated symmetric Pareto-optimal strategy profile.

We define the biased belief ψ_i^* as follows:

$$\psi_i^*(\alpha_1, \alpha_2 ..., \alpha_{99}, \alpha_{100}) = \left(\alpha_1, \alpha_2, ..., \frac{\alpha_{99}}{2}, \frac{\alpha_{99}}{2} + \alpha_{100}\right).$$

In what follows we show that $((\psi_1^*, \psi_2^*), (99, 99))$ is a strong BBE. Observe first that $\psi_i^*(99) = \left(0, ..., 0, \frac{1}{2}, \frac{1}{2}\right)$, which implies that $99 \in BR(\psi_i^*(99))$, and, thus, $(99, 99) \in NE\left(G_{(\psi_1^*, \psi_2^*)}\right)$. Let ψ_1' be an arbitrary perception bias of player i. Observe that player i never plays action 100 in a any Nash equilibrium of any biased game, because action 100 is not a best reply against any strategy of player j. Next observe that player i can obtain a payoff higher than 99 only if (1) player j chooses action 99 with a positive probability, and (2) player i chooses action 98 with a probability strictly higher than his probability of playing action 100. However, the biased belief ψ_j^* of player j implies that if player i chooses action 98 with a probability strictly higher than his probability of playing 100, then player j never chooses action 99 in any Nash equilibrium of the induced biased game because action 99 yields a strictly lower payoff to player j than action 98 against the perceived strategy of player i (because according to this perceived strategy, player i plays action 100 with a probability strictly less than player i's probability of playing either action 98 or action 99).

Note that the BBE equilibrium outcome $(99, 99)$ is consistent with level-1 behavior in the level-k and cognitive hierarchy literature (see, e.g., Stahl and Wilson, 1994; Nagel, 1995; Costa-Gomes, Crawford, and Broseta, 2001; Camerer, Ho, and Chong, 2004), according to which each agent believes that his opponent is following a focal non-strategic action (the action 100 in the traveler's dilemma), and best-replies to this belief. The notion of BBE can help explain why such level-k behavior induces a strategic advantage in the long run, and why, therefore, it is likely to emerge in an equilibrium.

A.5 Hawk-Dove Game

The following example characterizes the set of BBE (and their supporting beliefs) in a hawk-dove game (which is a game of strategic substitutes).

Example 12 (*The Hawk-dove game*)**.** The hawk-dove (or "chicken") game is a two-action game in which each player i has two actions: d_i (interpreted as a "dove"-like action of willingness to share a resource with the opponent) and h_i (interpreted as a "hawk"-like action of insistence on getting the whole resource, even if this requires fighting against the opponent), and where the ordinal preferences of each player i are (h_i, d_j) (getting the resource) $\succ (d_i, d_j)$ (sharing the resource) $\succ (d_i, h_j)$ (not getting the the resource) $\succ (h_i, h_j)$ (being involved in a serious fight). Table 6 presents the payoff of a general two-action hawk-dove game, where we have normalized, without loss of generality, the payoff of each player when playing action profile (d_i, d_j) $((h_i, h_j))$ to be one (zero), and where each g_i positive and each l_i is in the interval $(0, 1)$.

Table 6: Hawk-Dove Game ($g_1, g_2 > 0$ and $l_1, l_2 \in (0, 1)$)

	d_2	h_2
d_1	$1, 1$	$1 - l_1, 1 + g_2$
h_1	$1 + g_1, 1 - l_1$	$0, 0$

It is well known that the hawk-dove game admits three equilibria: two pure equilibria (d_1, h_2) and (h_1, d_2), and one mixed equilibrium (α_1^*, α_2^*), where the probability that player i plays action

α_i^* is

$$\alpha_i^* = \frac{1-l_j}{g_j+(1-l_j)} \in (0,1), \quad \text{and} \quad \pi\left(\alpha_i^*, \alpha_j^*\right) = \alpha_j^* \cdot (1+g_i) = 1 - \frac{g_i}{g_i+(1-l_i)} \cdot l_i.$$

The undominated minmax payoff of each player coincides with the minmax payoff of each player (as there are no dominated actions), and it is equal to $M_i^U = 1 - l_i$, which is obtained when the opponent plays h_j.

Applying the analysis of the previous section shows that the game admits 3 classes of BBE:

- Pure equilibrium hawk-dove: $\left(\left(\psi_i^*, \psi_j^*\right), (0,1)\right)$, where (1) ψ_i^* is an arbitrary monotone biased belief that satisfies $\psi_i^*(0) \geq \alpha_i^*$, and (2) ψ_j^* is an arbitrary monotone biased belief that satisfies $\psi_i^*(0) \leq \alpha_i^*$.

- Mixing (with less weight to playing dove), wishful thinking, and one-directional blindness: $\left(\left(\psi_1^*, \psi_2^*\right), (\beta_1, \beta_2)\right)$, where for each player i: (1) the payoff is above the minmax payoff: $\pi_i(\beta_i, \beta_j) \geq 1 - l_i$, (2) $\beta_i \in (0, \alpha_i^*)$ (i.e., agents play dove less often in the unique Nash equilibrium), (3) wishful thinking: $\psi_i^*(\beta_j) = \alpha_j^* > \beta_j$, and (4) responsiveness only to good news: $\psi_i^*(\alpha) = \alpha_j^*$ for each $\alpha \leq \beta_j$, and $\psi_i^*(\alpha) > \alpha_j^*$ for each $\alpha > \beta_j$.

Observe that any profile (β_1, β_2) where $\beta_i \in (\alpha_i^*, 1)$ cannot be a BBE outcome. If $\beta_j = 1$, then player i can gain by deviating to $\psi_i' \equiv 1$, as the unique equilibrium of the new biased game is $(0_i, 1_j)$, which induces a higher payoff to player i relative to (β_i, β_j). If $\beta_j < 1$, then player j can gain by deviating into $\psi_j' \equiv 1$, as the only possible equilibria of the new biased game are $(1_i, 0_j)$ and $(\beta_i, 1_j)$, both of which induce a higher payoff to player j relative to (β_i, β_j).

B Evolutionary Interpretation of BBE

In this section we present a formal definition of strong BBE that is exactly analogous to the definition of a stable configuration à la Dekel, Ely, and Yilankaya (2007). This shows that our static solution concept of strong BBE captures evolutionary stability in the same way as the solution concepts used in the literature on "indirect evolution of preferences." Finally, we illustrate a detailed example of a possible learning dynamic that may result in convergence to strong BBE.

B.1 Evolutionary Definition of Strong BBE à la Dekel, Ely, and Yilankaya (2007)

In this subsection we present a definition of a strong BBE that is completely analogous to the definition of a stable configuration a la Dekel, Ely, and Yilankaya (2007) (henceforth DEY) for the case of perfect observability of the opponent's type (i.e., $p = 1$ in DEY).

In the adaptation of the notion of stable configuration à la Dekel, Ely, and Yilankaya (2007) to our setup we change two aspects (and only these aspects):

1. We deal with general two-player games played between two different populations, rather than DEY's setup that deals with symmetric two-player games played within a single population.

2. Each agent in DEY's model is endowed with a type that determines the agent's subjective preferences. By contrast, in our setup each agent is endowed with a type that determines the agent's monotone biased belief.

3. We focus on homogeneous configurations. DEY's general definitions allow one to deal with heterogeneous configurations (in which different incumbents may have different types). However, their results mainly deal with homogeneous configurations (in which all incumbents have the same type). Therefore, to ease notation, we focus on homogeneous configurations in our adaptation of DEY's definitions.

After adapting DEY's definition of a homogeneous configuration (page 689 in DEY) to the three aspects mentioned above, their definition is as follows:

Definition 13. A (homogeneous) *configuration* is a pair $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$, where, for each player i, function ψ_i^* is a monotone biased belief of player i and s_i^* is a strategy of player i satisfying $s_i^* \in BR\left(\psi_i^*\left(s_j^*\right)\right)$.

It is immediate that any monotone weak BBE is a configuration.

Next, DEY present a notion of a balanced configuration (page 689 in DEY) that is trivially satisfied by any homogeneous configuration.

Consider two continuum populations of mass one that follow a configuration $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$. Assume that one of these populations (say, population i) is invaded by a small group of $0 < \epsilon << 1$ mutants with a different biased belief $\psi_i' \neq \psi_i^*$. DEY assume that (1) such a mutation can destabilize a configuration by resulting in the mutants achieving a higher fitness than the incumbents of the same population[1] i, and (2) the incumbents continue to play the same behavior among themselves (what DEY calls "focal equilibria").

Let Ψ_i be the set of all biased beliefs of player i. Following DEY (page 690 in DEY) we define $N_{i,\epsilon}\left(\psi_i^*, \psi_i'\right) \in \Delta\left(\Psi_i\right)$ to be the set of distributions over biased beliefs in population i resulting from entry by no more than ϵ mutants. Formally,

$$N_{i,\epsilon}\left(\psi_i^*, \psi_i'\right) = \left\{\mu_i' \in \Delta\left(\Psi_i\right) | \mu_i' = \left(1-\epsilon'\right) \cdot \psi_i^* + \epsilon' \cdot \psi_i',\ \epsilon' < \epsilon\right\}.$$

Given a configuration $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ and a post-entry distribution of biased beliefs in population i $\tilde{\mu}_i \in N_{i,\epsilon}\left(\psi_i^*, \psi_i'\right)$, a *post-entry focal configuration* is a pair $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$, where (1) $s_i' \in BR\left(\psi_i'\left(s_j'\right)\right)$ is interpreted as the mutant's strategy, and (2) $s_j' \in BR\left(\psi_j^*\left(s_i'\right)\right)$ is interpreted as population j's strategy against the mutants. The incumbents are assumed to play the same pre-entry strategies $\left(s_i^*, s_j^*\right)$ when being matched among themselves. Let $B\left(\tilde{\mu}_i\right)$ denote the set of all post-entry focal configurations.

Following DEY (Definition 3 on page 691 in DEY), we define DEY-stability of a configuration as follows.

[1] Under imperfect observability, a mutant can destabilize a configuration by unraveling the original equilibrium behavior, thereby causing the incumbents' strategies to substantially diverge following the mutant's entry into the population. This cannot happen under perfect observability, as the incumbents can always exhibit the same equilibrium behavior when being matched against other incumbents (see, page 690 in DEY for a discussion of focal equilibria).

Definition 14. Configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is *DEY-stable* if there exists $\epsilon > 0$ such that for every player i, every biased belief ψ_i', every post-entry distribution of biased beliefs $\tilde{\mu}_i \in N_{i,\epsilon}(\psi_i^*, \psi_i')$, and every post-entry focal configuration $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$, the mutants are weakly outperformed relative to the incumbents' payoff (in their own population), i.e., $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

B.2 Equivalence between the Definitions

The following result shows that the definition of a stable configuration coincides with our definition of strong BBE.

Proposition 10. *A configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is DEY-stable iff it is a strong BBE.*

Proof. "If" part: Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a strong BBE. Let $\epsilon > 0$, $i \in \{1, 2\}$, and $\psi_i' \in \Psi_i$. Let $\tilde{\mu}_i \in N_{i,\epsilon}\left(\psi_j^*, \psi_i'\right)$ be a post-entry distribution of biased beliefs. Let $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ be a post-entry focal configuration. The fact that $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ is a post-entry focal configuration implies that $s_i' \in BR\left(\psi_i'\left(s_j'\right)\right)$ and $s_j' \in BR\left(\psi_j^*\left(s_i'\right)\right)$. The fact that it is a strong BBE implies that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$, which shows that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is DEY-stable.

"Only if" part: Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be DEY-stable configuration. Let $i \in \{1, 2\}$ and $\psi_i' \in \Psi_i$. Let $\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$ be an equilibrium of the new biased game. Let $\epsilon > 0$. Let $\tilde{\mu}_i \in N_{i,\epsilon}(\psi_i^*, \psi_i')$ be a post-entry distribution of biased beliefs. For each $\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$, let $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ be a post-entry focal configuration. The assumption that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is DEY-stable implies that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$. This implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a strong BBE. □

Remark (Allowing multiple simultaneous invasions of mutants). The definition of DEY-stability presented above is unaffected when various groups of mutants simultaneously invade one of the populations. By contrast, if one were to require a stable configuration to resist simultaneous invasions of two groups of mutants, one invasion of each population, it would require a refinement of the concept of strong BBE, in the spirit of Maynard-Smith and Price's (1973) notion of evolutionary stability, such that if both ψ_1' and ψ_2' are best replies against configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$, then (1) ψ_1^* should be a strictly better reply against ψ_2' (relative to ψ_1'), and (2) ψ_2^* should be a strictly better reply against ψ_1' (relative to ψ_2').

Similarly, one can formulate a definition of stability equivalent to that of monotone BBE by requiring the mutants to be weakly outperformed in at least one post-entry focal configuration.

Definition 15. Configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is *weakly stable* if there exists $\epsilon > 0$ such that for every player i, every biased belief ψ_i', and every post-entry distribution of biased beliefs $\tilde{\mu}_i \in N_{i,\epsilon}(\psi_i^*, \psi_i')$, there exists a post-entry focal configuration $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ in which the mutants are weakly outperformed relative to the incumbents' payoff, i.e., $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

The following result shows that the definition of a weakly stable configuration coincides with our definition of weak BBE. The simple proof, which is analogous to the proof of 10, is omitted for brevity.

Proposition 11. *A configuration* $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ *is weakly stable iff it is a monotone weak BBE.*

Finally, one can formulate a definition of stability equivalent to that of a BBE by requiring the mutants to be weakly outperformed in at least one plausible post-entry focal configuration.

Definition 16. Given configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$, $\epsilon > 0$, $i \in \{1, 2\}$, biased belief ψ_i', and a post-entry distribution of biased beliefs $\tilde{\mu}_i \in N_{i,\epsilon}(\psi_i^*, \psi_i')$, we say that a post-entry focal configuration $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ is *implausible* if: (1) $\psi_j^*(s_i') = \psi_j^*(s_i^*)$, (2) $s_j' \neq s_j^*$, and (3) $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j^*\right)\right)$ is a post-entry focal configuration. A post-entry focal configuration is *plausible* if it is not implausible.

Definition 17. Configuration $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is *plausibly stable* if there exists $\epsilon > 0$ such that for every player i, every biased belief ψ_i', and every post-entry distribution of biased beliefs $\tilde{\mu}_i \in N_{i,\epsilon}(\psi_i^*, \psi_i')$, there exists a plausible post-entry focal configuration $\left(\left(\tilde{\mu}_i, \psi_j^*\right), \left(s_i', s_j'\right)\right)$ in which the mutants are weakly outperformed relative to the incumbents' payoff, i.e., $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

The following result shows that the definition of a plausibly stable configuration coincides with our definition of BBE. The simple proof, which is analogous to the proof of 10, is omitted for brevity.

Proposition 12. *A configuration* $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ *is* plausibly *stable iff it is a BBE.*

B.3 Illustration of the Evolutionary Interpretation

Similar to DEY, we have presented a reduced-form static notion of evolutionary stability, without formally modeling a detailed dynamics according to which the biased beliefs and the strategies co-evolve. In Section 3.6 we present the essential features of this evolutionary process, which are analogous to DEY's essential features (see first paragraph in Section 2.2 in DEY): agents are endowed by biased beliefs, these biased-beliefs induce equilibrium behavior in the biased game (presumably by a relatively quick adjustment of the biased players that leads to equilibrium behavior in the biased game), behavior determines "success," and success (the material payoffs) regulates the evolution of biased beliefs (presumably by a slow process in which agents occasionally die and are replaced by new agents who are more likely to mimic the biased beliefs of more successful incumbents).

In what follows, we illustrate this evolutionary process and its underlying dynamics in an example. Specifically, we present a strong BBE in an "input" game and we illustrate how this strong BBE can persist, given plausible evolutionary dynamics through which the composition of the population evolves.

Example 13 (*Example 9 revisited*)**.** Consider the following "input" game. Let $S_i = S_j = [0, 1]$, and let the payoff function be $\pi_i(s_i, s_j, \rho) = s_i \cdot s_j - \frac{s_i^2}{2\rho}$, where the parameter $\frac{1}{\rho}$ is interpreted as the cost of effort, and we assume that $\rho \in (0.5, 1)$. One can show that (1) the best-reply function of each agent is to exert an effort that is ρ times smaller than the opponent's (i.e., $BR(s_j) = \rho \cdot s_j$), (2) in the unique Nash equilibrium of the unbiased game each player exerts no effort $s_i = s_j = 0$, and (3) the strategy profile (ρ, ρ) yields a payoff of $\rho^2 - \frac{\rho}{2} > 0$, which is the highest symmetric payoff among all strategy profiles in which agents do not use strictly dominated strategies. Let ψ_i^* be the following biased-belief function:

$$\psi_i^*(s_j) = \begin{cases} \frac{s_j}{\rho} & s_j < \rho \\ 1 & s_j \geq \rho. \end{cases}$$

In Example 9 we have shown that $((\psi_1^*, \psi_2^*), (\rho, \rho))$ is a strong BBE. In what follows we illustrate how this strong BBE can persist. Consider a small group of mutants of population i who have undistorted beliefs. Assume that, initially, the incumbents of population j use the same strategy against the mutants as they use against the incumbents of population i (i.e., strategy ρ), and the mutants gradually learn to best reply to the incumbents' behavior by playing ρ^2. Recall that we assume that the agents of population j identify the mutants as a separate group of agents who behave differently than the rest of population j (without assuming that the incumbents of population j know anything about the biased beliefs of the mutants). These incumbents perceive the mutants' play as ρ (due to the incumbents' biased beliefs), and gradually learn to best reply to this perceived strategy by playing ρ^2. This, in turn, induces the mutants to adapt their play to playing ρ^3, and, in response, the incumbents of population j adapt their play against the mutants and play ρ^3 (the best reply to the mutants' perceived strategy ρ^2). This mutual gradual adaptation process continues until the play in the matches between incumbents of population j and mutants of population i converges to $(0, 0)$.

Finally, following the convergence of the behavior in the matches against the mutants to $(0, 0)$, a slow flow of new agents begins to influence the composition of the population. Each new agent randomly chooses a mentor among the agents in his own population, where agents with higher fitness are more likely to be chosen as mentors. As the mutants get a much lower payoff (0) than the incumbents of population i ($\rho^2 - \frac{\rho}{2} > 0$) in the underlying game, their fitness is expected to be lower, and they are much less likely to be chosen as mentors. As a result the share of mutants in the population slowly shrinks until they disappear from the population.

C Principal-Agent (Subgame-Perfect) Definition of BBE

In this appendix we present an equivalent definition of BBE as a subgame-perfect equilibrium of a two-stage game in which in the first round each player chooses the biased belief of the agent who will play on his behalf in the second round.

C.1 The Two-Stage Game Γ_G

Given an underlying two-player normal-form game $G = (S, \pi)$ define Γ_G as the following four-player two-stage extensive-form game. The four players in the game Γ are: principal 1 and principal 2 (who choose representative agents for the second stage), agent 1 (who plays on behalf of principal 1 in round 2), and agent 2 (who plays on behalf of principal 2 in round 2).

The game Γ_G has 2 stages. In the first stage, the principals simultaneously choose biased beliefs for their agents. That is, each principal i chooses a biased belief $\psi_i : S_j \to S_j$ for agent i. In the second stage the agents simultaneously choose their strategies. That is, each agent i chooses strategy $s_i \in S_i$. The payoff of each principal i is $\pi_i(s_i, s_j)$. The payoff of each agent i is $\pi_i(\psi_i(s_i), s_j)$. Let Ψ_i be the set of all feasible (monotone) biased beliefs of agent i.

A pure strategy profile of Γ_G (henceforth Γ_G-strategy profile) is a tuple $(\psi_1, \psi_2, \sigma_1, \sigma_2)$, where each ψ_i is a biased belief, and each $\sigma_i : \Psi_1 \times \Psi_2 \to S_i$ is a function assigning a strategy to each pair of (monotone) biased beliefs. Let $SPE(\Gamma_G)$ denote the set of all subgame-perfect equilibria of Γ.

C.2 Subgame-Perfect Definition of Weak BBE

The following result shows that a weak BBE is equivalent to a subgame-perfect equilibrium of Γ. Formally:

Proposition 13. *Let G be a game. Strategy profile $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a weak BBE of G iff there exists a subgame-perfect equilibrium $\left(\left(\psi_1^*, \psi_2^*\right), \left(\sigma_1^*, \sigma_2^*\right)\right)$ of Γ_G satisfying $\sigma_i^*\left(\psi_i^*\right) = s_i^*$ for each player i.*

Proof. "If side": Let $\left(\left(\psi_1^*, \psi_2^*\right), \left(\sigma_1^*, \sigma_2^*\right)\right) \in SPE\left(\Gamma_G\right)$ be a subgame-perfect equilibrium of Γ satisfying $\sigma_i^*\left(\psi_i^*\right) = s_i^*$ for each player i. Let ψ_i' be a biased belief of player i. Let $s_1' = \sigma_1^*\left(\psi_i', \psi_j^*\right)$ and $s_2' = \sigma_2^*\left(\psi_i', \psi_j^*\right)$. The fact that $\left(\left(\psi_1^*, \psi_2^*\right), \left(\sigma_1^*, \sigma_2^*\right)\right) \in SPE\left(\Gamma_G\right)$ implies that (1) $\left(s_1', s_2'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ and (2) $\pi_i\left(s_1', s_2'\right) \leq \pi_i\left(s_1^*, s_2^*\right)$. This implies that $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a weak BBE of G.

"Only if side": Let $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ be a weak BBE of G. We define $\left(\sigma_1^*, \sigma_2^*\right)$ as follows:[2] (1) $\sigma_i^*\left(\psi_1^*, \psi_2^*\right) = s_i^*$, (2) for each biased belief $\psi_i' \neq \psi_i^*$, define $\sigma_i^*\left(\psi_i', \psi_j^*\right) = s_i'$ and $\sigma_j^*\left(\psi_i', \psi_j^*\right) = s_j'$ such that $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ and $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$ (such a pair $\left(s_i', s_j'\right)$ exists due to $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ being a weak BBE of G), and (3) for each pair of biased beliefs $\psi_i' \neq \psi_i^*$ and $\psi_j' \neq \psi_j^*$, define $\sigma_i^*\left(\psi_i', \psi_j'\right) = s_i'$ and $\sigma_j^*\left(\psi_i', \psi_j'\right) = s_j'$ such that $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j'\right)}\right)$. The definition of $\left(\sigma_1^*, \sigma_2^*\right)$ immediately implies that $\left(\left(\psi_1^*, \psi_2^*\right), \left(\sigma_1^*, \sigma_2^*\right)\right) \in SPE\left(\Gamma_G\right)$. □

C.3 Subgame-Perfect Definition of BBE

Next, we present an equivalent definition of a BBE as a refinement of a subgame-perfect equilibrium of Γ_G. Specifically, a subgame-perfect equilibrium $\left(\psi_1^*, \psi_2^*, \sigma_1^*, \sigma_2^*\right)$ is required to remain a subgame-perfect equilibrium even after changing the off-the equilibrium path behavior to a different Nash equilibrium of the induced subgame in which (I) a single player (say, player j) has deviated to a different biased-belief, (II) the non-deviator perceives the deviator's strategy in the same way as the original on-the-equilibrium path opponent's strategy, and (III) the non-deviator changes his behavior such that after the change it coincides with his on-the-equilibrium path behavior. Formally,

Definition 18. A subgame-perfect equilibrium $\left(\psi_1^*, \psi_2^*, \sigma_1^*, \sigma_2^*\right) \in SPE\left(\Gamma_G\right)$ is a *plausible subgame-perfect equilibrium* if (I) the biased beliefs ψ_1^* and ψ_2^* are monotone, and (II) $\left(\psi_1^*, \psi_2^*, \sigma_1', \sigma_2'\right) \in SPE\left(\Gamma\right)$ for each pair of second-stage strategies σ_1', σ_2' satisfying: (1) $\left(\psi_1', \psi_2', \sigma_1', \sigma_2'\right) \in SPE\left(\Gamma_G\right)$ for some pair of first-stage strategies $\left(\psi_1', \psi_2'\right)$ (i.e., second-stage behavior is consistent with equilibrium behavior in all subgames) and (2) if $\sigma_i'\left(\psi_1', \psi_2'\right) \neq \sigma_i^*\left(\psi_1', \psi_2'\right)$, then: (I) $\psi_i' = \psi_i^*$ and $\psi_j' \neq \psi_j^*$, (II) $\psi_i^*\left(\sigma_j'\left(\psi_1', \psi_2'\right)\right) = \psi_i^*\left(\sigma_j^*\left(\psi_1', \psi_2'\right)\right)$, and (III) $\sigma_j'\left(\psi_1', \psi_2'\right) = \sigma_j^*\left(\psi_1', \psi_2'\right)$.

Proposition 14. *Let G be a game. Strategy profile $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a BBE of G iff there exists a plausible subgame-perfect equilibrium $\left(\left(\psi_1^*, \psi_2^*\right), \left(\sigma_1^*, \sigma_2^*\right)\right)$ of Γ_G satisfying $\sigma_i^*\left(\psi_i^*\right) = s_i^*$ for each player i.*

The simple proof, which is analogous to the proof of Proposition 13, is omitted for brevity.

[2] The definition of $\left(\sigma_1^*, \sigma_2^*\right)$ relies on the axiom of choice.

C.4 Subgame-Perfect Definition of Strong BBE

Finally, we present an equivalent definition of a strong BBE as a refinement of a subgame-perfect equilibrium of Γ, which remains an equilibrium even after changing off the equilibrium path in subgames to other Nash equilibria of the induced subgames. Formally,

Definition 19. A subgame-perfect equilibrium $(\psi_1^*, \psi_2^*, \sigma_1^*, \sigma_2^*) \in SPE(\Gamma_G)$ is a *strong subgame-perfect equilibrium* if (I) the biased beliefs ψ_1^* and ψ_2^* are monotone, and (II) $(\psi_1^*, \psi_2^*, \sigma_1', \sigma_2') \in SPE(\Gamma_G)$ for each pair of second-stage strategies σ_1', σ_2' satisfying: (1) $\left(\psi_1', \psi_2', \sigma_1', \sigma_2'\right) \in SPE(\Gamma_G)$ for some pair of first-stage strategies ψ_1', ψ_2' (i.e., second-stage behavior is consistent with equilibrium behavior in all subgames) and (2) $\sigma_i'(\psi_1^*, \psi_2^*) = \sigma_i^*(\psi_1^*, \psi_2^*)$ (i.e., behavior after (ψ_1^*, ψ_2^*) is unchanged).

Our final result shows that a strong BBE is equivalent to a strong subgame-perfect equilibrium of Γ. Formally:

Proposition 15. *Let G be a game. Strategy profile $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a strong BBE of G iff there exists a strong subgame-perfect equilibrium $((\psi_1^*, \psi_2^*), (\sigma_1^*, \sigma_2^*))$ of Γ satisfying $\sigma_i^*(\psi_i^*) = s_i^*$ for each player i.*

The simple proof, which is analogous to the proof of Proposition 13, is omitted for brevity.

D Discontinuous Biased Beliefs

In this appendix we present an alternative definition of BBE that relaxes the assumption that biased beliefs have to be continuous. We show that all BBE characterized in the main text remain BBE when deviators are allowed to use discontinuous biased beliefs.

D.1 Adapted Definitions: Quasi-equilibria

We redefine a *biased belief* $\psi_i : S_j \to S_j$ to be an arbitrary (rather than continuous) function that assigns to each strategy of the opponent a (possibly distorted) belief about the opponent's play. The definition of a configuration (ψ^*, s^*) is left unchanged (i.e., we require that $\left(s_i^*, s_j^*\right) \in NE(G_{\psi^*})$).

Recall that a configuration is a BBE if each biased belief is a best reply to the opponent's biased belief, in the sense that an agent who chooses a different biased belief is weakly outperformed in the induced equilibrium of the new biased game. Allowing discontinuous beliefs implies that some biased games $G_{(\psi_1,\psi_2)}$ in which one (or both) of the biases are discontinuous may not admit Nash equilibria. This requires us to adapt the definition of a BBE to deal with behavior in biased games that do not admit Nash equilibria. We do so by assuming that the resulting behavior in a biased game that does not admit a Nash equilibrium is a "j-quasi-equilibrium," in which the non-deviator (player j) best replies to the perceived behavior of the deviator (player i), while the deviator is allowed to play arbitrarily. Formally:

Definition 20. Let (ψ_i, ψ_j) be a profile of biased beliefs, and let j be one of the players (interpreted as the non-deviator); then we define $QE_j\left(G_{(\psi_i,\psi_j)}\right)$ as the set of j-quasi-equilibria of the biased

game $G_{(\psi_i,\psi_j)}$ as follows:

$$QE_j\left(G_{(\psi_i,\psi_j)}\right)=\begin{cases}NE\left(G_{(\psi_i,\psi_j)}\right) & NE\left(G_{(\psi_i,\psi_j)}\right)\neq\emptyset\\ \{(s_i,s_j)\,|s_j\in BR\left(\psi_j\left(s_i\right)\right)\} & NE\left(G_{(\psi_i,\psi_j)}\right)=\emptyset.\end{cases}$$

Note that any biased game admits a j-quasi-equilibrium.

D.2 Adapted Definitions: BBE′

We redefine our notions of BBE as follows, and write them as BBE′. In a strong BBE′, the deviator (player i) is required to be outperformed in all j-quasi-equilibria, and biased beliefs are required to be monotone. In a weak BBE′, the deviator is required to be outperformed in at least one j-quasi-equilibrium. The notion of a BBE′ is in between these two notions. Specifically, in a BBE′, the biased beliefs are required to be monotone, and, in addition, the deviator (player i) is required to be outperformed in at least one plausible j-quasi-equilibrium of the new biased game, where implausible j-quasi-equilibria are defined as follows. We say that a j-quasi-equilibrium of a biased game induced by a deviation of player i is implausible if (1) player i's strategy is perceived by the non-deviating player j as coinciding with player i's original strategy, (2) player j plays differently relative to his original strategy, and (3) if player j were playing his original strategy, this would induce a j-quasi-equilibrium of the biased game. That is, implausible j-quasi-equilibria are those in which the non-deviating player j plays differently against a deviator even though player j has no reason to do so: player j does not observe any change in player i's behavior, and player j's original behavior remains an equilibrium of the biased game. Formally:

Definition 21. Given configuration (ψ^*,s^*), deviating player i, and biased belief ψ_i', we say that a j-quasi-equilibrium of the biased game $\left(s_i',s_j'\right)\in QE_j\left(G_{(\psi_i',\psi_j^*)}\right)$ is *implausible* if: (1) $\psi_j^*\left(s_i'\right)=\psi_j^*\left(s^*{}_i\right)$, (2) $s_j^*\neq s_j'$, and (3) $\left(s_i',s_j^*\right)\in QE_j\left(G_{(\psi_i',\psi_j^*)}\right)$. A j-quasi-equilibrium is *plausible* if it is not implausible. Let $PQE_j\left(G_{(\psi_i',\psi_j^*)}\right)$ be the set of all plausible j-quasi-equilibria of the biased game $G_{(\psi_i',\psi_j^*)}$.

Note that it is immediate from Definition 21 and the nonemptiness of $QE_j\left(G_{(\psi_i',\psi_j^*)}\right)$ that $PQE_j\left(G_{(\psi_i',\psi_j^*)}\right)$ is nonempty.

Definition 22. Configuration (ψ^*,s^*) is:

1. *a strong BBE′* if (I) each biased belief ψ_i^* is monotone, and (II) $\pi_i\left(s_i',s_j'\right)\leq\pi_i\left(s_i^*,s_j^*\right)$ for every player i, every biased belief ψ_i', and every j-quasi-equilibrium $\left(s_i',s_j'\right)\in QE_j\left(G_{(\psi_i',\psi_j^*)}\right)$;
2. *a weak BBE′* if for every player i and every biased belief ψ_i', there exists a j-quasi-equilibrium $\left(s_i',s_j'\right)\in QE_j\left(G_{(\psi_i',\psi_j^*)}\right)$, such that $\pi_i\left(s_i',s_j'\right)\leq\pi_i\left(s_i^*,s_j^*\right)$;

3. a *BBE′* if (I) each biased belief ψ_i^* is monotone, and (II) for every player i and every biased belief ψ_i', there exists a plausible j-quasi-equilibrium $\left(s_i', s_j'\right) \in PQE_j\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$, such that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

It is immediate that any strong BBE′ is a BBE′, and that any BBE′ is a weak BBE′.

(resp., strong, weak) BBE′ (ψ^*, s^*) is continuous if each biased function ψ_i^* is continuous. Note, that deviators are allowed to choose discontinuous biased beliefs.

D.3 Robustness of BBE to Discontinuous Biased Beliefs

In what follows we observe that all the BBE that we characterize in all the results of the paper are also BBE′. That is, all of our BBE are robust to allowing deviators to use discontinuous biased beliefs. Specifically, any BBE (resp., weak BBE, strong BBE) that is characterized in any result (or example) in the paper, is a continuous BBE′ (resp., weak continuous BBE′, strong continuous BBE′).

The reason why this observation is true is that in all the arguments in the proofs of the paper's results for why a configuration $\left(\left(\psi_i^*, \psi_j^*\right), \left(s_i^*, s_j^*\right)\right)$ is a BBE, when we show that a deviator (player i) is outperformed after deviating to biased belief ψ_i' and after the players play strategy profile $\left(s_i', s_j'\right)$, we rely only on the assumption that the non-deviator (player j) best replies to the deviator (i.e., that $s_j \in BR\left(\psi_j^*\left(s_i'\right)\right)$, which is implied by assuming $\left(s_i', s_j'\right) \in QE_j\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$), and we do not use in any of the arguments the assumption that the deviator plays a best reply (i.e., we do not rely on $s_i \in BR\left(\psi_i'\left(s_j'\right)\right)$ in any of the proofs).

E Partial Observability

Throughout the paper we assume that if an agent deviates to a different biased belief, then the opponent always observes this deviation. In this appendix, we relax this assumption, and show that our results hold also in a setup with partial observability (some results hold for any level of partial observability, while others hold for a sufficiently high level of observability).

E.1 Restricted Biased Games

Let $p \in [0, 1]$ denote the probability that an agent who is matched with an opponent who deviates to a different biased belief *privately* observes the opponent's deviation (henceforth, *observation probability*). If an agent does not observe the deviation, then he continues playing his original configuration's strategy.

Our definitions of configuration and biased game remain unchanged. We now define a restricted biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$ as a game with a payoff function according to which (1) each player's payoff is determined by the opponent's perceived strategy, and (2) the non-deviator is restricted to playing s_j^* with probability p (i.e., when not observing the opponent's deviation). Formally:

Definition 23. Given an underlying game $G = (S, \pi)$, a profile of biased beliefs $\left(\psi_i', \psi_j^*\right)$, and a strategy s_j^* of player j (interpreted as the non-deviator), let the *restricted biased game* $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)} = \left(S, \tilde{\pi}\left(\psi_i', \psi_j^*, s_j^*, p\right)\right)$ be defined as follows:

$$\tilde{\pi}_i\left(\psi_i', \psi_j^*, s_j^*, p\right)(s_i, s_j) = p \cdot \pi_i\left(s_i, \psi_i'(s_j)\right) + (1-p) \cdot \pi_i\left(s_i, \psi_i'\left(s_j^*\right)\right), \text{ and}$$

$$\tilde{\pi}_j\left(\psi_i', \psi_j^*, s_j^*, p\right)(s_i, s_j) = p \cdot \pi_j\left(s_j, \psi_j^*(s_i)\right) + (1-p)\,\pi_i\left(s_j^*, \psi_j^*(s_i)\right).$$

A Nash equilibrium of a p-restricted biased game is defined in the standard way. Formally, a pair of strategies $s^* = (s_1', s_2')$ is a Nash equilibrium of a restricted biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$, if each s_i' is a best reply against the perceived strategy of the opponent, i.e.,

$$s_i' = argmax_{s_i \in S_i}\left(\tilde{\pi}_i\left(\psi_i', \psi_j^*, s_j^*, p\right)\left(s_i, s_j'\right)\right).$$

Let $NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*\right)}\right) \subseteq S_1 \times S_2$ denote the set of all Nash equilibria of the restricted biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$.

Observe that the set of strategies of a biased game is convex and compact, and the payoff function $\tilde{\pi}_i\left(\psi_i', \psi_j^*, s_j^*, p\right) : S_i \times S_j \to \mathbb{R}$ is weakly concave in the first parameter and continuous in both parameters. This implies (due to a standard application of Kakutani's fixed-point theorem) that each restricted biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$ admits a Nash equilibrium (i.e., $NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right) \neq \emptyset$).

E.2 p-BBE

We are now ready to define our equilibrium concept. Configuration (ψ^*, s^*) is a p-BBE if each biased belief is a best reply to the opponent's biased belief, in the sense that an agent who chooses a different biased belief is weakly outperformed in the induced equilibrium of the new restricted biased game. We present three versions of p-BBE that differ with respect to the equilibrium selection when the new biased game admits multiple equilibria. In a strong p-BBE (I) each biased-belief is monotone, and (II) the deviator is required to be outperformed in all Nash equilibria of the new restricted biased game. In a weak BBE, the deviator is required to be outperformed in at least one equilibrium of the new restricted biased game.

The notion of a p-BBE is in between these two notions. Specifically, in at p-BBE (I) each biased-belief is monotone, and (II) the deviator is required to be outperformed in at least one plausible equilibrium of the new restricted biased game, where implausible equilibria are defined as follows. We say that a Nash equilibrium of a restricted biased game induced by a deviation of player i is implausible if (1) player i's strategy is perceived by the non-deviating player j as coinciding with player i's original strategy, (2) player j plays differently relative to his original strategy, and (3) if player j were playing his original strategy, this would induce an equilibrium of the biased game. That is, implausible equilibria are those in which the non-deviating player j plays differently against a deviator even though player j has no reason to do so: player j does not observe any change in player i's behavior, and player j's original behavior remains an equilibrium of the biased game.

Formally:

Definition 24. Given configuration (ψ^*, s^*), deviating player i, and biased belief ψ_i', we say that a Nash equilibrium of the restricted biased game $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$ is *implausible* if: (1) $\psi_j^*\left(s_i'\right) = \psi_j^*\left(s^*{}_i\right)$, (2) $s_j^* \neq s_j'$, and (3) $\left(s_i', s_j^*\right) \in NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$. An equilibrium is *plausible* if it is not implausible. Let $PNE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$ be the set of all plausible equilibria of the biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$.

Note that it is immediate from Definition 24 and the nonemptiness of $NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$ that $PNE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$ is nonempty.

Definition 25. Configuration (ψ^*, s^*) is:

1. *a strong p-BBE* if (I) each biased belief ψ_i^* is monotone, and (II) $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$ for every player i, every biased belief ψ_i', and every Nash equilibrium $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$;
2. *a weak p-BBE* if for every player i and every biased belief ψ_i', there exists a Nash equilibrium $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$, such that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$;
3. a *p-BBE* if (I) each biased belief ψ_i^* is monotone, and (II) for every player i and every biased belief ψ_i', there exists a plausible Nash equilibrium $\left(s_i', s_j'\right) \in PNE\left(G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}\right)$, such that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

It is immediate that: (1) any strong p-BBE is a p-BBE, and that any p-BBE is a weak p-BBE, and (2) the definition of 1-BBE (resp., weak 1-BBE, strong 1-BBE) coincides with the original definition of BBE (resp., weak BBE, strong BBE).

E.3 Extension of Results

In what follows we sketch how to extend our results to the setup of partial observability. The adaptations of the proofs are relatively simple, and, for brevity, we only sketch the differences with respect to the original proofs.

E.3.1 Adaptation of Section 4 (Nash Equilibria and BBE Outcomes)

The example that some Nash equilibria cannot be supported as the outcomes of weak P-BBE with undistorted beliefs can be extended for any $p > 0$.

Example 14 (*Example 1 revisited. Cournot equilibrium cannot be supported by undistorted beliefs*)**.** Consider the following symmetric Cournot game with linear demand $G = (S, \pi)$: $S_i = [0, 1]$ and $\pi_i\left(s_i, s_j\right) = s_i \cdot (1 - s_i - s_j)$ for each player i. The unique Nash equilibrium of the game is $s_i^* = s_j^* = \frac{1}{3}$, which yields both players a payoff of $\frac{1}{9}$. Fix observation probability $p > 0$. Assume to the contrary that this outcome can be supported as a weak p-BBE by the undistorted beliefs $\psi_i^* = \psi_j^* = I_d$. Fix

a sufficiently small $0 < \epsilon << 1$. Consider a deviation of player 1 to the blind belief $\psi_i' \equiv \frac{1}{3} - 2 \cdot \epsilon$. Note that this blind belief has a unique best reply: $s_i' = \frac{1}{3} + \epsilon$. The unique equilibrium of the restricted biased game $G_{\left(\psi_i', \psi_j^*, s_j^*, p\right)}$ is $s_j' = \frac{1}{3} - \frac{\epsilon}{2}$, $s_i' = \frac{1}{3} + \epsilon$, which yields the deviator a payoff of $\frac{1}{9} + \frac{\epsilon}{6} - \frac{\epsilon^2}{2}$ with probability p (when his deviation is observed by player 2) and a payoff of $\frac{1}{9} - \epsilon^2$ with probability $1 - p$ (when his deviation is not observed by player 2). For a sufficiently small $\epsilon > 0$ the expected payoff of the deviator is strictly larger than $\frac{1}{9}$.

All the results of Section 4 hold for any observation probability $p \in [0, 1]$ with minor adaptations to the proofs.

Proposition 16 (Proposition 1 extended)**.** *Let (s_1^*, s_2^*) be a (strict) Nash equilibrium of the game $G = (S, \pi)$. Let $\psi_1^* \equiv s_2^*$ and $\psi_2^* \equiv s_1^*$. Then $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a (strong) p-BBE for any $p \in [0, 1]$.*

Claim 2 (Claim 1 extended)*.* The unique Nash equilibrium payoff of a zero-sum game is also the unique payoff in any weak p-BBE for any $p \in [0, 1]$.

Proposition 17 (Proposition 2 extended)**.** *If a game admits a strictly dominant strategy s_i^* for player i, then any weak p-BBE outcome is a Nash equilibrium of the underlying game.*

E.3.2 Adaptation of Section 6 (Main Results)

Adaptation of Subsection 6.1 (Preliminary Result) Minor adaptations of the proof of Proposition 3 show that it holds for any $p \in [0, 1]$. Formally:

Proposition 18. *Let $p \in [0, 1]$. If a strategy profile $s^* = (s_1^*, s_2^*)$ is a weak p-BBE outcome, then (1) the profile s^* is undominated and (2) $\pi_i(s^*) \geq M_i^U$.*

Adaptation of Subsection 6.2 (Games with Strategic Complements) Minor adaptations to the proofs of the results of Subsection 6.2 show that most of these results (namely, part (1) of Proposition 4 and Corollaries 2 and 3) hold for any $p \in [0, 1]$, while part (2) of Proposition 4 holds for p-s sufficiently close to one. Formally:

Proposition 19 (*Proposition* 4 extended)**.** *Let G be a game with* strategic substitutes and positive externalities.

1. *Fix $p \in [0, 1]$. Let (s_1^*, s_2^*) be a p-BBE outcome. Then (s_1^*, s_2^*) is (I) undominated, and for each player i: (II) $\pi_i\left(s_i^*, s_j^*\right) \geq M_i^U$, and (III) $s_i^* \leq \max\left(BR\left(s_j^*\right)\right)$ (underinvestment).*

2. *Let (s_1^*, s_2^*) be an undominated profile satisfying for each player i: (II') $\pi_i\left(s_i^*, s_j^*\right) > \tilde{M}_i^U$, and (III) $s_i^* \leq \max\left(BR\left(s_j^*\right)\right)$. Then there exists $\bar{p} < 1$ such that (s_1^*, s_2^*) is a p-BBE outcome for any $p \in [\bar{p}, 1]$.*
   Moreover, if $\pi_i(s_i, s_j)$ is strictly concave then (s_1^*, s_2^*) is a strong p-BBE outcome for any $p \in [\bar{p}, 1]$.

Corollary 7. *Fix $p \in [0, 1]$. Let G be a game with strategic complements and positive externalities with a lowest Nash equilibrium $(\underline{s}_1, \underline{s}_2)$ satisfying $\underline{s}_1 < \max(S_i)$ for each player i. Let (s_1^*, s_2^*) be a p-BBE outcome. Then $\underline{s}_i \leq s_i^*$ for each player i.*

Corollary 8. *Fix $p \in [0,1]$. Let G be a game with positive externalities and strategic complements. Let*
$((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a p-BBE. If $s_i^* \notin \{\min(S_i), \max(S_i)\}$, then player i exhibits wishful thinking (i.e., $\psi_i^*\left(s_j^*\right) \geq s_j^*$).

One can also adapt the examples of Section 6.2 (and, similarly, the examples of Sections 6.3 and 6.4) to sufficiently high ps.

Adaptation of Section 6.3 (Games With Strategic Substitutes) Minor adaptations to the proofs of the results of Section 6.3 show that most of these results (namely, part (1) of Proposition 5 and Corollaries 4 and 5) hold for any $p \in [0,1]$, while part (2) of Proposition 5 holds for p-s sufficiently close to one. Formally:

Proposition 20 (*Proposition* 5 extended)**.** *Let G be a game with strategic substitutes and positive externalities.*

1. *Fix $p \in [0,1]$. Let (s_1^*, s_2^*) be a p-BBE outcome. Then (s_1^*, s_2^*) is (I) undominated, and for each player i: (II) $\pi_i\left(s_i^*, s_j^*\right) \geq M_i^U$, and (III) $s_i^* \geq \min\left(BR\left(s_j^*\right)\right)$ (overinvestment).*

2. *Let (s_1^*, s_2^*) be an undominated profile satisfying for each player i: (II') $\pi_i\left(s_i^*, s_j^*\right) > \tilde{M}_i^U$, and (III) $s_i^* \geq \min\left(BR\left(s_j^*\right)\right)$. Then there exists $\bar{p} < 1$ such that (s_1^*, s_2^*) is a p-BBE outcome for any $p \in [\bar{p}, 1]$.*
   Moreover, if $\pi_i(s_i, s_j)$ is strictly concave then (s_1^*, s_2^*) is a strong p-BBE outcome for any $p \in [\bar{p}, 1]$.

Corollary 9. *Fix $p \in [0,1]$. Let G be a game with strategic substitutes and positive externalities. Let (s_1^*, s_2^*) be a BBE outcome. Then, there exists a Nash equilibrium of the underlying game (s_1^e, s_2^e), and a player i such that $s_i^e \geq s_i^*$.*

Corollary 10. *Fix $p \in [0,1]$. Let G be a game with strategic substitutes and positive externalities. Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a p-BBE. If $s_i^* \notin \{\min(S_i), \max(S_i)\}$,* then player i exhibits wishful thinking (i.e., $\psi_i^*\left(s_j^*\right) \geq s_j^*$).

Adaptation of Section 6.3 (Games With Strategic Opposites) Minor adaptations to the proofs of the results of Section 6.3 show that most of these results (namely, part (1) of Proposition 6, and Corollary 6) hold for any $p \in [0,1]$, while part (2) of Proposition 6 holds for p-s sufficiently close to one. Formally:

Proposition 21. *Let G be a game with positive externalities and strategic opposites*: $\frac{\partial \pi_1(s_1,s_2)}{\partial s_1} > 0$ and $\frac{\partial \pi_2(s_1,s_2)}{\partial s_1} < 0$ for each pair of strategies s_1, s_2.

1. *Fix $p \in [0,1]$. Let (s_1^*, s_2^*) be a p-BBE outcome. Then (s_1^*, s_2^*) is (I) undominated: (II) $\pi_i\left(s_i^*, s_j^*\right) \geq M_i^U$ for each player i, and (III) $s_1^* \leq \max(BR(s_2^*))$ and $s_2^* \geq \min(BR(s_1^*))$ (underinvestment of player 1 and overinvestment of player 2).*

2. *Let* (s_1^*, s_2^*) *be a profile satisfying: (I) undominated, (II)* $\pi_i\left(s_i^*, s_j^*\right) > \tilde{M}_i^U$ *for each player* i*, and (III)* $s_1^* \leq \max\left(BR\left(s_2^*\right)\right)$ *and* $s_2^* \geq \min\left(BR\left(s_1^*\right)\right)$*. Then there exists* $\bar{p} < 1$ *such that* (s_1^*, s_2^*) *is a p-BBE outcome for any* $p \in [\bar{p}, 1]$*.*

Corollary 11. *Fix* $p \in [0,1]$*. Let* $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ *be a p-BBE of a game with* positive externalities *and strategic opposites (i.e.,* $\frac{\partial \pi_1(s_1,s_2)}{\partial s_1} > 0$ and $\frac{\partial \pi_2(s_1,s_2)}{\partial s_1} < 0$ for each pair of strategies s_1, s_2*). If* $s_i^* \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$*,* then player i exhibits pessimism (i.e., $\psi_i^*\left(s_j^*\right) \leq s_j^*$).

E.3.3 Adaptation of Section 7 (Additional Results)

Adaptation of Section 7.1 (BBE with Strategic Stubbornness) In what follows we show how to extend Example 5 to the setup of partial observability (while we leave the extension of the general result, Proposition 7, to future research). The example focuses on Cournot competition. We show that for each level of partial observability $p \in [0,1]$, there exists a strong BBE in which one of the players (1) has a blind belief and (2) plays a strategy that is between the Nash equilibrium strategy and the Stackelberg strategy (and the closer it is to the Stackelberg strategy, the higher the value of p), while the opponent has undistorted beliefs. The first player's (resp., opponent's) payoff is strictly increasing (resp., decreasing) in p: it converges to the Nash equilibrium payoff when $p \to 0$, and it converges to the Stackelberg leader's (resp., follower's) payoff when $p \to 1$.

Example 15 (*Example 5 revisited*)**.** Consider the symmetric Cournot game with linear demand: $G = (S, \pi)$: $S_i = \mathbb{R}^+$ and $\pi_i\left(s_i, s_j\right) = s_i \cdot (1 - s_i - s_j)$ for each player i. Let $p \in [0,1]$ be the observation probability. Then

$$\left(\left(\frac{1-p}{3-p}, I_d\right), \left(\frac{1}{3-p}, \frac{2-p}{2\cdot(3-p)}\right)\right)$$

is a strong BBE that yields a payoff of $\frac{2-p}{2\cdot(3-p)}$ to player 1, and yields a payoff of $\left(\frac{2-p}{2\cdot(3-p)}\right)^2$ to player 2. Observe that player 1's payoff is increasing in p, and it converges to the Nash equilibrium (resp., Stackelberg leader's) payoff of $\frac{1}{9}$ ($\frac{1}{8}$) when $p \to 0$ ($p \to 1$). Further observe that player 2's payoff is decreasing in p, and it converges to the Nash equilibrium (resp., Stackelberg follower's) payoff of $\frac{1}{9}$ ($\frac{1}{16}$) when $p \to 0$ ($p \to 1$). The argument that $\left(\left(\frac{1-p}{3-p}, I_d\right), \left(\frac{1}{3-p}, \frac{2-p}{2\cdot(3-p)}\right)\right)$ is a strong BBE is sketched as follows: (1) $\left\{\left(\frac{1}{3-p}, \frac{2-p}{2\cdot(3-p)}\right)\right\} = NE\left(G_{\left(\frac{1-p}{3-p}, I_d\right)}\right)$ (because $\frac{1}{3-p}$ is the unique best reply against $\frac{1-p}{3-p}$ and $\frac{2-p}{2\cdot(3-p)}$ is the unique best reply against $\frac{1}{3-p}$) ; (2) for any biased belief ψ_2', player 1 keeps playing $\frac{1}{3-p}$ due to having a blind belief, and as a result player 2's payoff is at most $\left(\frac{2-p}{2\cdot(3-p)}\right)^2$; and (3) for any biased belief ψ_1' inducing a deviating player 1 to play strategy x, player 2 plays $\frac{1-x}{2}$ (the unique best reply against x) with probability p (when observing the deviation), and player 2 plays $\frac{2-p}{2\cdot(3-p)}$ (the original configuration strategy) with the remaining probability of $1-p$. Thus, the payoff of a deviating player 1 who deviates into playing strategy x is

$$\pi\left(x\right) := p\cdot x\cdot\left(\frac{1-x}{2}\right) + (1-p)\cdot x\cdot\left(1 - x - \frac{2-p}{2\cdot(3-p)}\right) = \left(1 - \frac{p}{2}\right)\cdot x\cdot(1-x) - \frac{(2-p)\cdot(1-p)}{2\cdot(3-p)}\cdot x,$$

where this payoff function $\pi\left(x\right)$ is strictly concave in x with a unique maximum at $x = \frac{1}{3-p}$ (the

unique solution to the FOC $0 = \frac{\partial \pi}{\partial x} = (1 - \frac{p}{2}) \cdot (1 - 2 \cdot x) - \frac{(2-p)\cdot(1-p)}{2\cdot(3-p)})$.

Extending the Folk Theorem Results for Sufficiently High p-s The main results of Section 7.2, show folk theorem results for: (1) monotone BBE in games that admit best replies with full undominated support, and (2) strong BBE in interval games with a payoff function that is strictly concave in the agent's strategy, and weakly convex in the opponent's strategy. Minor adaptations of each proof can show that each result can be extended to p-s that are sufficiently close to one. Formally:

Proposition 22 (*Proposition* 8 extended)**.** *Let G be a finite game that admits best replies with full undominated support. Let (s_1^*, s_2^*) be an undominated strategy profile that induces for each player a payoff above his minmax payoff (i.e., $\pi_i(s_1^*, s_2^*) > M_i^U \ \forall i \in \{1,2\}$). Then there exists $\bar{p} < 1$, such that (s_1^*, s_2^*) is a monotone weak p-BBE outcome for each $p \in [\bar{p}, 1]$.*

Proposition 23 (*Proposition* 9 extended)**.** *Let $G = (S, \pi)$ be an interval game. Assume that for each player i, $\pi_i(s_i, s_j)$ is strictly concave in s_i and weakly convex in s_j. If (s_1^*, s_2^*) is undominated and $\pi_i(s_1^*, s_2^*) > M_i^U$ for each player i, then there exists $\bar{p} < 1$, such that (s_1^*, s_2^*) is a strong p-BBE outcome for each $p \in [\bar{p}, 1]$.*

Sketch of adapting the proofs of Propositions 22 and 23 to the setup of partial observability. Observe that the gain of an agent who deviates to a different biased belief, when his deviation is unobserved by the opponent, is bounded (due to the payoff of the underlying game being bounded). When the deviation is observed by the opponent, the agent is strictly outperformed, given the BBE constructed in the proofs of Propositions 8 and 9. This implies that there exists $\bar{p} < 1$ sufficiently close to one, such that the loss of a mutant when being observed by his opponent outweighs the mutant's gain when being unobserved for any $p \in [\bar{p}, 1]$. □

F Proofs

F.1 Proof of Proposition 4

Part 1: Proposition 3 implies (I) and (II). It remains to show (III, overinvestment). Let $\left(\left(\psi_i^*, \psi_j^*\right), \left(s_i^*, s_j^*\right)\right)$ be a BBE. Assume to the contrary that $s_i^* < \min\left(BR\left(s_j^*\right)\right)$. Consider a deviation of player i to a blind belief that the opponent always plays strategy s_j^*(i.e., $\psi_i' \equiv s_j^*$). Let $\left(s_i', s_j'\right) \in PNE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ be a plausible equilibrium of the new biased game. Observe first that $s_i' \in BR\left(\psi_i'\left(s_j'\right)\right) = BR\left(s_j^*\right)$. This implies that $s_i' > s_i^*$, and, thus, due to the monotonicity of ψ_j^* we have: $\psi_j^*(s_i') \geq \psi_j^*(s_i^*)$. We consider two cases:

1. If $\psi_j^*(s_i') > \psi_j^*(s_i^*)$, then the strategic complementarity implies that $s_j' \geq \min\left(BR\left(\psi_j^*(s_i')\right)\right) \geq \max\left(BR\left(\psi_j^*(s_i^*)\right)\right) \geq s_j^*$, and this, in turn, implies that player i strictly gains from his deviation: $\pi_i\left(s_i', s_j'\right) \geq \pi_i\left(s_i', s_j^*\right) > \pi_i\left(s_i^*, s_j^*\right)$, a contradiction.

2. If $\psi_j^*(s_i') = \psi_j^*(s_i^*)$, then $\left(s_i', s_j^*\right) \in PNE\left(G_{(\psi_i', \psi_j^*)}\right)$ and $\pi_i\left(s_i', s_j^*\right) > \pi_i\left(s_i^*, s_j^*\right)$, which contradicts that $\left(\left(\psi_i^*, \psi_j^*\right), \left(s_i^*, s_j^*\right)\right)$ is a BBE.

Part 2: Assume that strategy profile (s_1^*, s_2^*) satisfies I, II, and III. For each player i let $s_i^e = \min\left(BR^{-1}(s_i^*)\right)$. For every player i and every strategy $s_i < s_i^*$ define $X(s_i)$ as the set of strategies s_i' for which player i is worse off (relative to $\pi_i(s_1^*, s_2^*)$) if he plays strategy s_i, while player j plays a best-reply to s_i'. Formally:

$$X_{s^*}(s_i) = \left\{s_i' \in S_i | \pi_i(s_i, s_j) \leq \pi_i\left(s_i^*, s_j^*\right) \ \forall s_j \in BR(s_i')\right\}.$$

The assumption that $\pi_i\left(s_i^*, s_j^*\right) > \tilde{M}_i^U$ implies that $X_{s^*}(s_i)$ is nonempty for each s_i. The assumption of strategic complements implies that $X_{s^*}(s_i)$ is an interval starting at $\min(S_i)$. Let $\phi_{s^*}(s_i) = \sup(X_{s^*}(s_i))$. The assumption that the payoff function is continuously twice differentiable implies that $\phi_{s^*}(s_i)$ is continuous. The assumption that $s_j^e = \min\left(BR^{-1}(s_i^*)\right)$ implies that $\lim_{s_i \nearrow s_i^*}(\phi_{s^*}(s_i)) = s_i^e$. These observations imply that for each player j there exists a monotone biased belief ψ_j^* satisfying (1) $\psi_j^*(s_i) = s_i^e$ for each $s_i \geq s_i^*$ and (2) $\psi_j^*(s_i) \leq \phi_{s^*}(s_i)$ for each $s_i < s_i^*$ with an equality only if $\phi_{s^*}(s_i) = \min(S_i)$.

We now show that these properties of (ψ_1^*, ψ_2^*) imply that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE (a strong BBE if $\pi_i(s_i, s_j)$ is strictly concave in s_i). Consider a deviation of player i into an arbitrary biased belief ψ_i'. For each $s_i' \geq s_i^*$, and each $\left(s_i', s_j'\right) \in PNE\left(G_{(\psi_i', \psi_j^*)}\right)$ ($\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$), the fact that $\psi_j^*(s_i') = \psi_j^*(s_i^*)$ implies that $s_j' = s_j^*$, and due to assumption (III) of overinvestment and the concavity of the payoff function: $\pi_i\left(s_i', s_j'\right) = \pi_i\left(s_i', s_j^*\right) \leq \pi_i\left(s_i^*, s_j^*\right)$. For each $s_i' < s_i^*$, and each $\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$, the fact that $\psi_j^*(s_i') \leq \phi_{s^*}(s_i')$ with an equality only if $\phi_{s^*}(s_i') = \min(S_i)$ (and, thus, $\psi_j^*(s_i') \in X_{s^*}(s_i')$) implies that $\pi_i\left(s_i', s_j'\right) \leq \pi_i(s_1^*, s_2^*)$. This shows that player i cannot gain from his deviation, which implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a (strong) BBE.

F.2 Proof of a Lemma Required for Corollary 2

Lemma 1. *Let G be a game with strategic complements and positive externalities with a lowest Nash equilibrium $(\underline{s}_1, \underline{s}_2)$ satisfying $\underline{s}_i < \max(S_i)_i$ for each player i. Let $s_1^* < \underline{s}_1$. Then for each $s_2^* \in S_2$ either (1) $s_1^* < \min(BR(s_2^*))$ or (2) $s_2^* < \min(BR(s_1^*))$.*

Proof. Assume first that $s_2^* > \underline{s}_2$. The fact that $\underline{s}_1 \in BR(\underline{s}_2)$ and $s_2^* > \underline{s}_2$, together with the strategic complements, imply that $s_1^* < \underline{s}_1 < \min(BR(s_2^*))$. We are left with the case where $s_2^* \leq \underline{s}_2$. Consider a restricted game in which the set of strategies of each player i is restricted to being strategies that are at most s_i^*. The game is a game of strategic complements, and, thus, it admits a pure Nash equilibrium $\left(s_i', s_j'\right)$. The minimality of $(\underline{s}_1, \underline{s}_2)$ implies that $\left(s_i', s_j'\right)$ cannot be a Nash equilibrium of the unrestricted game. The strategic complements and the concavity of the payoff jointly imply that if $\left(s_i', s_j'\right)$ is not a Nash equilibrium of the unrestricted game, then there is player i for which $s_i^* = s_i' < \min\left(BR\left(s_j'\right)\right) \leq \min\left(BR\left(s_j^*\right)\right)$. □

F.3 Proof of a Lemma Required for Corollary 3

Lemma 2. *Let G be a game with positive externalities and strategic complementarity of the payoff of player i (i.e., $\frac{\partial^2 \pi_i(s_i,s_j)}{\partial s_i \cdot \partial s_j} > 0$ for each s_i, s_j). Then $s'_j < s_j$ implies that $\max\left(BR\left(s'_j\right)\right) \leq \min\left(BR\left(s_j\right)\right)$ with an equality only if $\max\left(BR\left(s'_j\right)\right) = \min\left(BR\left(s_j\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\}$.*

Proof. The inequality $s'_j < s_j$ and the strategic complementarity of the payoff of player i implies that $\frac{\partial \pi_i\left(s_i,s'_j\right)}{\partial s_i} < \frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$ for each $s_i \in S_i$, which implies that whenever $\max\left(BR\left(s'_j\right)\right) \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$, then

$$\begin{aligned}\max\left(BR\left(s'_j\right)\right) &= \max\left\{s_i^* \in S_i | \left.\frac{\partial \pi_i\left(s_i, s'_j\right)}{\partial s_i} = 0\right|_{s_i=s_i^*}\right\} \\ &< \min\left\{s_i^* \in S_i | \left.\frac{\partial \pi_i\left(s_i, s_j\right)}{\partial s_i} = 0\right|_{s_i=s_i^*}\right\} \leq \min\left(BR\left(s_j\right)\right).\end{aligned}$$

This shows that the strict inequality holds whenever $\max\left(BR\left(s'_j\right)\right) \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$. It remains to show that the weak inequality (namely, $\max\left(BR\left(s'_j\right)\right) \leq \min\left(BR\left(s_j\right)\right)$) holds when $\max\left(BR\left(s'_j\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\}$. If $\max\left(BR\left(s'_j\right)\right) = \min\left(S_i\right)$ then this is immediate. Assume that $\max\left(BR\left(s'_j\right)\right) = \max\left(S_i\right)$. Then:

$$\begin{aligned}\max\left(S_i\right) &= \max\left\{s_i^* \in S_i | \left.\frac{\partial \pi_i\left(s_i, s'_j\right)}{\partial s_i} \geq 0\right|_{s_i=s_i^*}\right\}. \\ &\leq \min\left\{s_i^* \in S_i | \left.\frac{\partial \pi_i\left(s_i, s_j\right)}{\partial s_i} \geq 0\right|_{s_i=s_i^*}\right\} \leq \min\left(BR\left(s_j\right)\right).\end{aligned}$$

□

F.4 Proof of Proposition 5

The proof is analogous to the proof of Proposition 4, and is presented for completeness.

Part 1: Proposition 3 implies (I) and (II). It remains to show (III) (underinvestment). Let $\left(\left(\psi_i^*, \psi_j^*\right), \left(s_i^*, s_j^*\right)\right)$ be a BBE. Assume to the contrary that $s_i^* > \max\left(BR\left(s_j^*\right)\right)$. Consider a deviation of player i to a blind belief that the opponent always plays strategy s_j^*(i.e., $\psi'_i \equiv s_j^*$). Let $\left(s'_i, s'_j\right) \in PNE\left(G_{\left(\psi'_i, \psi_j^*\right)}\right)$ be a plausible equilibrium of the new biased game. Observe first that $s'_i \in BR\left(\psi'_i\left(s'_j\right)\right) = BR\left(s_j^*\right)$. This implies that $s'_i < s_i^*$, and, thus, due to the monotonicity of ψ_j^* we have: $\psi_j^*\left(s'_i\right) \leq \psi_j^*\left(s_i^*\right)$. We consider two cases:

1. If $\psi_j^*\left(s'_i\right) < \psi_j^*\left(s_i^*\right)$, then the strategic substitutability implies that $s'_j \geq \min\left(BR\left(\psi_j^*\left(s'_i\right)\right)\right) \geq \max\left(BR\left(\psi_j^*\left(s_i^*\right)\right)\right) \geq s_j^*$, and this, in turn, implies that player i strictly gains from his deviation: $\pi_i\left(s'_i, s'_j\right) \geq \pi_i\left(s'_i, s_j^*\right) > \pi_i\left(s_i^*, s_j^*\right)$, a contradiction.

2. If $\psi_j^*(s_i') = \psi_j^*(s_i^*)$, then $\left(s_i', s_j^*\right) \in PNE\left(G_{(\psi_i', \psi_j^*)}\right)$ and $\pi_i\left(s_i', s_j^*\right) > \pi_i\left(s_i^*, s_j^*\right)$, which contradicts that $\left(\left(\psi_i^*, \psi_j^*\right), \left(s_i^*, s_j^*\right)\right)$ is a BBE.

Part 2: Assume that strategy profile (s_1^*, s_2^*) satisfies I, II, and III. For each player i let $s_i^e = \max\left(BR^{-1}(s_i^*)\right)$. For each player i and each strategy $s_i > s_i^*$ define $X(s_i)$ as the set of strategies s_i' for which player i is worse off (relative to $\pi_i(s_1^*, s_2^*)$) if he plays strategy s_i, while player j plays a best-reply to s_i'. Formally:

$$X_{s^*}(s_i) = \left\{s_i' \in S_i | \pi_i(s_i, s_j) \leq \pi_i\left(s_i^*, s_j^*\right) \ \forall s_j \in BR(s_i')\right\}.$$

The assumption that $\pi_i\left(s_i^*, s_j^*\right) > \tilde{M}_i^U$ implies that $X_{s^*}(s_i)$ is nonempty for each s_i. The assumption of strategic substitutes implies that $X_{s^*}(s_i)$ is an interval ending at $\max(S_i)$. Let $\phi_{s^*}(s_i) = \inf(X_{s^*}(s_i))$. The assumption that the payoff function is continuously twice differentiable implies that $\phi_{s^*}(s_i)$ is continuous. The assumption that $s_j^e = \max\left(BR^{-1}(s_i^*)\right)$ implies that $\lim_{s_i \searrow s_i^*}(\phi_{s^*}(s_i)) = s_i^e$. These observations imply that for each player j there exists a monotone biased belief ψ_j^* satisfying (1) $\psi_j^*(s_i) = s_i^e$ for each $s_i \leq s_i^*$ and (2) $\psi_j^*(s_i) \geq \phi_{s^*}(s_i)$ for each $s_i > s_i^*$ with an equality only if $\phi_{s^*}(s_i) = \max(S_i)$.

We now show that these properties of (ψ_1^*, ψ_2^*) imply that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE (a strong BBE if $\pi_i(s_i, s_j)$ is strictly concave in s_i). Consider a deviation of player i to an arbitrary biased belief ψ_i'. For each $s_i' \leq s_i^*$, and each $\left(s_i', s_j'\right) \in PNE\left(G_{(\psi_i', \psi_j^*)}\right)$ ($\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$), the fact that $\psi_j^*(s_i') = \psi_j^*(s_i^*)$ implies that $s_j' = s_j^*$ and, due to assumption (III) of underinvestment and the concavity of the payoff function, it follows that $\pi_i\left(s_i', s_j'\right) = \pi_i\left(s_i', s_j^*\right) \leq \pi_i\left(s_i^*, s_j^*\right)$. For each $s_i' > s_i^*$, and each $\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^*)}\right)$, the fact that $\psi_j^*(s_i') \geq \phi_{s^*}(s_i')$ with an equality only if $\phi_{s^*}(s_i) = \max(S_i)$ (and, thus, $\psi_j^*(s_i') \in X_{s^*}(s_i')$) implies that $\pi_i\left(s_i', s_j'\right) \leq \pi_i(s_1^*, s_2^*)$. This shows that player i cannot gain from his deviation, which implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a (strong) BBE.

F.5 Proof of a Lemma Required for Corollary 4

Lemma 3. *Let G be a game with strategic substitutes and positive externalities. Let (s_1^*, s_2^*) be a strategy profile satisfying $s_i^* > s_i^e$ for each player i and each Nash equilibrium $(s_1^e, s_2^e) \in NE(G)$. Then, either (1) $s_1^* > \max(BR(s_2^*))$ or (2) $s_2^* > \max(BR(s_1^*))$.*

Proof. Consider a restricted game in which the set of strategies of each player i is restricted to being strategies that are at least s_i^*. The restricted game is a game with strategic substitutes, and, thus, it admits a pure Nash equilibrium (s_1', s_2') (recall, that after relabeling the set of strategies of one of the players, the game becomes supermodular, and because of this the game admits a pure Nash equilibrium due to Milgrom and Roberts, 1990). The assumption that $s_i^* > s_i^e$ for each player i and each Nash equilibrium $(s_1^e, s_2^e) \in NE(G)$ implies that (s_1', s_2') cannot be a Nash equilibrium of the unrestricted game. The concavity of the payoff and the strategic substitutes jointly imply that if $\left(s_i', s_j'\right)$ is not a Nash equilibrium of the unrestricted game, then there is a player i for which $s_i^* = s_i' > \max\left(BR\left(s_j'\right)\right) \geq \max\left(BR\left(s_j^*\right)\right)$. □

F.6 Proof of Corollary 5

The proof is analogous to Corollary 3, and is presented for completeness. Assume to the contrary that $\psi_i^*\left(s_j^*\right) < s_j^*$. Lemma 4 (below) implies that $\min\left(BR\left(\psi_i^*\left(s_j^*\right)\right)\right) \geq \max\left(BR\left(s_j^*\right)\right)$ with an equality only if

$$\min\left(BR\left(\psi_i^*\left(s_j^*\right)\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\}.$$

Part 1 of Proposition 5 and the definition of a monotone BBE imply that

$$\min\left(BR\left(\psi_i^*\left(s_j^*\right)\right)\right) \leq s_i^* \leq \max\left(BR\left(s_j^*\right)\right).$$

The previous inequalities jointly imply that

$$\min\left(BR\left(\psi_i^*\left(s_j^*\right)\right)\right) = s_i^* = \max\left(BR\left(s_j^*\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\},$$

which contradicts the assumption that $s_i^* \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$.

Lemma 4. *Let G be a game with positive externalities and strategic substitutability of the payoff of player i (i.e., $\frac{\partial^2 \pi_i(s_i,s_j)}{\partial s_i \cdot \partial s_j} > 0$ for each s_i, s_j). Then $s_j' < s_j$ implies that $\min\left(BR\left(s_j'\right)\right) \geq \max\left(BR\left(s_j\right)\right)$ with an equality only if $\min\left(BR\left(s_j'\right)\right) = \min\left(BR\left(s_j\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\}$.*

Proof. The proof is analogous to the proof of Lemma 1, and is presented for completeness. The inequality $s_j' < s_j$ and the strategic substitutability of the payoff of player i implies that $\frac{\partial \pi_i\left(s_i,s_j'\right)}{\partial s_i} > \frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$ for each $s_i \in S_i$, which implies that whenever $\min\left(BR\left(s_j'\right)\right) \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$, then

$$\begin{aligned}\min\left(BR\left(s_j'\right)\right) &= \min\left\{s_i^* \in S_i \middle| \frac{\partial \pi_i\left(s_i, s_j'\right)}{\partial s_i} = 0\middle|_{s_i=s_i^*}\right\} \\ &> \max\left\{s_i^* \in S_i \middle| \frac{\partial \pi_i\left(s_i, s_j\right)}{\partial s_i} = 0\middle|_{s_i=s_i^*}\right\} = \max\left(BR\left(s_j\right)\right).\end{aligned}$$

This shows that the strict inequality holds whenever $\min\left(BR\left(s_j'\right)\right) \notin \{\min\left(S_i\right), \max\left(S_i\right)\}$. It remains to show that the weak inequality (namely, $\min\left(BR\left(s_j'\right)\right) \geq \max\left(BR\left(s_j\right)\right)$) holds when $\min\left(BR\left(s_j'\right)\right) \in \{\min\left(S_i\right), \max\left(S_i\right)\}$. If $\min\left(BR\left(s_j'\right)\right) = \max\left(S_i\right)$ then this is immediate. Assume that $\min\left(BR\left(s_j'\right)\right) = \min\left(S_i\right)$. Then:

$$\begin{aligned}\min\left(S_i\right) &= \min\left\{s_i^* \in S_i \middle| \frac{\partial \pi_i\left(s_i, s_j'\right)}{\partial s_i} \geq 0\middle|_{s_i=s_i^*}\right\} \\ &\geq \max\left\{s_i^* \in S_i \middle| \frac{\partial \pi_i\left(s_i, s_j\right)}{\partial s_i} \geq 0\middle|_{s_i=s_i^*}\right\} \geq \max\left(BR\left(s_j\right)\right).\end{aligned}$$

□

F.7 Proof of Proposition 6

The proof is analogous to the proof of Proposition 4, and is presented for completeness.

Part 1: Proposition 3 implies (I) and (II). It remains to show (III). Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a BBE. We begin by showing overinvestment of player 2. Assume to the contrary that $s_2^* < \min\left(BR\left(s_j^*\right)\right)$. Consider a deviation of player 2 to a blind belief that the opponent always plays strategy s_1^* (i.e., $\psi_2' \equiv s_1^*$). Let $(s_1', s_2') \in PNE\left(G_{\left(\psi_1^*, \psi_2'\right)}\right)$ be a plausible equilibrium of the new biased game. Observe first that $s_2' \in BR\left(\psi_2'\left(s_1'\right)\right) = BR\left(s_1^*\right)$. This implies that $s_2' > s_2^*$, and, thus, due to the monotonicity of ψ_1^*, we have : $\psi_1^*\left(s_2'\right) \geq \psi_1^*\left(s_2^*\right)$. We consider two cases:

1. If $\psi_1^*\left(s_2'\right) > \psi_1^*\left(s_2^*\right)$, then the strategic complementarity of player 1's payoff implies that $s_1' \geq \min\left(BR\left(\psi_1^*\left(s_2'\right)\right)\right) \geq \max\left(BR\left(\psi_1^*\left(s_2^*\right)\right)\right) \geq s_1^*$, and, this, in turn, implies that player 2 strictly gains from his deviation: $\pi_2\left(s_1', s_2'\right) \geq \pi_2\left(s_1', s_2^*\right) > \pi_2\left(s_1^*, s_2^*\right)$, a contradiction.

2. If $\psi_1^*\left(s_2'\right) = \psi_1^*\left(s_2^*\right)$, then $\left(s_1^*, s_2'\right) \in PNE\left(G_{\left(\psi_1^*, \psi_2'\right)}\right)$ and $\pi_2\left(s_1^*, s_2'\right) > \pi_2\left(s_1^*, s_2^*\right)$, which contradicts that $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a BBE.

Next we show underinvestment of player 1. Assume to the contrary that $s_1^* > \max\left(BR\left(s_2^*\right)\right)$. Consider a deviation of player 1 to a blind belief that the opponent always plays strategy s_2^* (i.e., $\psi_1' \equiv s_2^*$). Let $(s_1', s_2') \in PNE\left(G_{\left(\psi_1', \psi_2^*\right)}\right)$ be a plausible equilibrium of the new biased game. Observe first that $s_1' \in BR\left(\psi_1'\left(s_2'\right)\right) = BR\left(s_2^*\right)$. This implies that $s_1' < s_2^*$ and, thus, due to the monotonicity of ψ_2^*, we have: $\psi_2^*\left(s_1'\right) \leq \psi_2^*\left(s_1^*\right)$. We consider two cases:

1. If $\psi_2^*\left(s_1'\right) < \psi_2^*\left(s_1^*\right)$, then the strategic substitutability of player 2's payoff implies that $s_2' \geq \min\left(BR\left(\psi_2^*\left(s_1'\right)\right)\right) \geq \max\left(BR\left(\psi_2^*\left(s_1^*\right)\right)\right) \geq s_2^*$, and this, in turn, implies that player 1 strictly gains from his deviation: $\pi_1\left(s_1', s_2'\right) \geq \pi_1\left(s_1', s_2^*\right) > \pi_1\left(s_1^*, s_2^*\right)$, a contradiction.

2. If $\psi_2^*\left(s_1'\right) = \psi_2^*\left(s_1^*\right)$, then $\left(s_1', s_2^*\right) \in PNE\left(G_{\left(\psi_1', \psi_2^*\right)}\right)$ and $\pi_1\left(s_1', s_2^*\right) > \pi_1\left(s_1^*, s_2^*\right)$, which contradicts that $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ is a BBE.

Part 2: Assume that strategy profile (s_1^*, s_2^*) satisfies I, II, and III. Let $s_1^e = \min\left(BR^{-1}\left(s_2^*\right)\right)$. For each strategy $s_2 < s_2^*$ define $X\left(s_2\right)$ as the set of strategies s_2' for which player 2 is worse off (relative to $\pi_2\left(s_1^*, s_2^*\right)$) if he plays strategy s_2, while player 1 plays a best-reply to s_2'. Formally:

$$X_{s^*}\left(s_2\right) = \left\{s_2' \in S_2 | \pi_2\left(s_1, s_2\right) \leq \pi_2\left(s_i^*, s_j^*\right) \ \forall s_1 \in BR\left(s_2'\right)\right\}.$$

The assumption that $\pi_2\left(s_1^*, s_2^*\right) > \tilde{M}_2^U$ implies that $X_{s^*}\left(s_2\right)$ is nonempty for each $s_2 \in S_2$. The assumption of strategic complements of player 1's payoff implies that $X_{s^*}\left(s_2\right)$ is an interval starting at $\min\left(S_2\right)$. Let $\phi_{s^*}\left(s_2\right) = \sup\left(X_{s^*}\left(s_2\right)\right)$. The assumption that the payoff function is continuously twice differentiable implies that $\phi_{s^*}\left(s_2\right)$ is continuous. The assumption that $s_1^e = \min\left(BR^{-1}\left(s_2^*\right)\right)$ implies that $\lim_{s_2 \nearrow s_2^*}\left(\phi_{s^*}\left(s_2\right)\right) = s_2^e$. These observations imply that there exists a monotone biased belief ψ_1^* satisfying (1) $\psi_1^*\left(s_2\right) = s_1^e$ and (2) $\psi_1^*\left(s_2\right) \leq \phi_{s^*}\left(s_2\right)$ for each $s_2 < s_2^*$ with an equality only if $\phi_{s^*}\left(s_2\right) = \min\left(S_2\right)$.

Let $s_2^e = \max\left(BR^{-1}\left(s_1^*\right)\right)$. For each strategy $s_1 > s_1^*$ define $X\left(s_1\right)$ as the set of strategies $s_1' \in S_1$ for which player 1 is worse off (relative to $\pi_2\left(s_1^*, s_2^*\right)$) if he plays strategy s_1, while player 2

plays a best-reply to s_1'. Formally:

$$X_{s^*}(s_2) = \left\{ s_1' \in S_1 | \pi_2(s_1, s_2) \leq \pi_2(s_1^*, s_2^*) \ \forall s_1 \in BR(s_2') \right\}.$$

The assumption that $\pi_1(s_1^*, s_2^*) > \tilde{M}_1^U$ implies that $X_{s^*}(s_1)$ is nonempty for each $s_1 \in S_1$. The assumption of strategic substitutes of player 2's payoff implies that $X_{s^*}(s_1)$ is an interval ending at $\max(S_1)$. Let $\phi_{s^*}(s_1) = \inf(X_{s^*}(s_1))$. The assumption that the payoff function is continuously twice differentiable implies that $\phi_{s^*}(s_1)$ is continuous. The assumption that $s_2^e = \max\left(BR^{-1}(s_1^*)\right)$ implies that $\lim_{s_1 \searrow s_1^*}(\phi_{s^*}(s_1)) = s_1^e$. These observations imply that there exists a monotone biased belief ψ_2^* satisfying (1) $\psi_2^*(s_1) = s_1^e$ and (2) $\psi_2^*(s_1) \geq \phi_{s^*}(s_1)$ for each $s_1 > s_1^*$ with an equality only if $\phi_{s^*}(s_1) = \max(S_1)$.

We now show that these properties of (ψ_1^*, ψ_2^*) imply that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE. Consider a deviation of player 2 into an arbitrary biased belief ψ_2'. For each $s_2' \geq s_2^*$, and each $(s_1', s_2') \in PNE\left(G_{(\psi_1^*, \psi_2')}\right)$, the fact that $\psi_1^*(s_2') = \psi_1^*(s_2^*)$ implies that $s_1' = s_1^*$, and due to assumption (III) of the overinvestment of player 2 and the concavity of the payoff function, we have $\pi_2(s_1', s_2') = \pi_2(s_1', s_2^*) \leq \pi_2(s_1^*, s_2^*)$. For each $s_2' < s_2^*$, and each $(s_1', s_2') \in NE\left(G_{(\psi_1^*, \psi_2')}\right)$, the fact that $\psi_1^*(s_2') \leq \phi_{s^*}(s_2')$ with an equality only if $\phi_{s^*}(s_2') = \min(S_2)$ implies that $\pi_2(s_1', s_2') \leq \pi_2(s_1^*, s_2^*)$. This shows that player 2 cannot gain from his deviation.

Finally, consider a deviation of player 1 to an arbitrary biased belief ψ_1'. For each $s_1' \leq s_1^*$, and each $(s_1', s_2') \in PNE\left(G_{(\psi_1', \psi_2^*)}\right)$, the fact that $\psi_2^*(s_1') = \psi_2^*(s_1^*)$ implies that $s_2' = s_2^*$, and due to assumption (III) of the underinvestment of player 1 and the concavity of the payoff function, we have $\pi_1(s_1', s_1') = \pi_1(s_1', s_3^*) \leq \pi_1(s_1^*, s_3^*)$. For each $s_1' > s_1^*$, and each $(s_1', s_3') \in NE\left(G_{(\psi_1', \psi_2^*)}\right)$, the fact that $\psi_2^*(s_1') \geq \phi_{s^*}(s_1')$ with an equality only if $\phi_{s^*}(s_1) = \max(S_1)$ implies that $\pi_1(s_1', s_2') \leq \pi_1(s_1^*, s_2^*)$. This shows that player 1 cannot gain from his deviation, which implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE.

F.8 Proof of Corollary 6 (Pessimism in Games with Strategic Opposites)

The proof is analogous to the proof of Corollary 3, and is presented for completeness.

Assume to the contrary that $\psi_i^*\left(s_j^*\right) > s_j^*$ for some player i. Assume first that $\psi_2^*(s_1^*) > s_1^*$; then Lemma 4 implies that $\max(BR(\psi_2^*(s_1^*))) \leq \min(BR(s_1^*))$ with an equality only if

$$\max(BR(\psi_2^*(s_1^*))) \in \{\min(S_2), \max(S_2)\}.$$

Part 1 of Proposition 6 and the definition of a monotone BBE imply that

$$\max(BR(\psi_2^*(s_1^*))) \geq s_2^* \geq \min(BR(s_1^*)).$$

The previous inequalities jointly imply that

$$\max(BR(\psi_2^*(s_1^*))) = s_2^* = \min(BR(s_1^*)) \in \{\min(S_2), \max(S_2)\},$$

which contradicts the assumption that $s_2^* \notin \{\min(S_2), \max(S_2)\}$.

We are left with the case of $\psi_1^*(s_2^*) > s_2^*$; then Lemma 2 implies that $\min(BR(\psi_1^*(s_2^*))) \geq \max(BR(s_2^*))$ with an equality only if

$$\min(BR(\psi_1^*(s_2^*))) \in \{\min(S_1), \max(S_1)\}.$$

Part 1 of Proposition 6 and the definition of a monotone BBE imply that

$$\min(BR(\psi_1^*(s_2^*))) \leq s_1^* \leq \max(BR(s_2^*)).$$

The previous inequalities jointly imply that

$$\min(BR(\psi_1^*(s_2^*))) = s_1^* = \max(BR(s_2^*)). \in \{\min(S_1), \max(S_1)\},$$

which contradicts the assumption that $s_1^* \notin \{\min(S_1), \max(S_1)\}$.

F.9 Proof of Proposition 9

Recall that we assume the payoff function π_i to be continuously twice differentiable. This implies that π_i is Lipschitz continuous. Let $K_i > 0$ be the Lipschitz constant of the payoff function π_i with respect to its first parameter, i.e., K_i satisfies

$$\|\pi_i(s_1, s_2) - \pi_i(s_1', s_2)\| \leq K_i \cdot \|s_1 - s_1'\|.$$

Assume that (s_1^*, s_2^*) is undominated and $\pi_i(s_1^*, s_2^*) > M_i^U$ for each player i. Let $0 < D_i = \pi_i(s_1^*, s_2^*) - M_i^U$. For each player j, let s_j^p be an undominated strategy that guarantees that player i obtains, at most, his minmax payoff M_i^U, i.e., $s_j^p = \operatorname{argmin}_{s_j \in S_j^U}(\max_{s_i \in S_i} \pi_i(s_i, s_j))$. The strict concavity of $\pi_i(s_i, s_j)$ with respect to s_i implies that the best-reply correspondence is a continuous one-to-one function. Thus, $BR^{-1}(s_i)$ is a singleton for each s_i, and we identify $BR^{-1}(s_i)$ with the unique element in this singleton set.

Let $\epsilon > 0$ be a sufficiently small number satisfying $\epsilon < \min\left(\frac{D_i}{K_i}, \frac{D_j}{K_j}\right)$. For each $\delta \in [0,1]$ define for each player i:

$$s_i^\delta = \frac{\epsilon - \delta}{\epsilon} \cdot s_i^* + \frac{\delta}{\epsilon} \cdot s_i^p.$$

Let ψ_i^ϵ be defined as follows:

$$\psi_i^\epsilon\left(s_j'\right) = \begin{cases} BR^{-1}\left(s_i^{\left|s_j' - s_j\right|}\right) & \left|s_j' - s_j\right| < \epsilon \\ BR^{-1}(s_i^p) & \left|s_j' - s_j\right| \geq \epsilon. \end{cases}$$

Note that ψ_i^ϵ is continuous. We now show that $((\psi_1^\epsilon, \psi_2^\epsilon), (s_1^*, s_2^*))$ is a strong BBE. Observe first that the definition of $(\psi_1^\epsilon, \psi_2^\epsilon)$ immediately implies that $(s_1^*, s_2^*) \in NE\left(G_{(\psi_1^\epsilon, \psi_2^\epsilon)}\right)$. Next, consider a deviation of player i to an arbitrary biased belief ψ_i'. Consider any equilibrium of the new biased game $\left(s_i', s_j'\right) \in NE\left(G_{(\psi_i', \psi_j^\epsilon)}\right)$. If $|s_i' - s_i| \geq \epsilon$, then the definition of $\psi_j^\epsilon(s_i')$ implies that $s_j^p = s_j'$, and that player i achieves a payoff of at most $M_i^U < \pi_i(s_1^*, s_2^*)$. If $s_i' = s_i^*$, then it is immediate that

$s'_j = s^*_j$ and that player i does not gain from his deviation. If $0 < |s'_i - s_i| < \epsilon$, then the definition of $\psi^\epsilon_j(s'_i)$ implies that

$$\pi_i\left(s'_i, s'_j\right) = \pi_i\left(s'_i, s_j^{|s'_i - s_i|}\right) = \pi_i\left(s'_i, \frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot s^*_j + \frac{|s'_i - s_i|}{\epsilon} \cdot s^p_j\right) \leq$$

$$\frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot \pi_i\left(s'_i, s^*_j\right) + \frac{|s'_i - s_i|}{\epsilon} \cdot \pi_i\left(s'_i, s^p_j\right) \leq \frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot \pi_i\left(s'_i, s^*_j\right) + \frac{|s'_i - s_i|}{\epsilon} \cdot M^U_i \leq$$

$$\frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot \pi_i\left({s^*}_i, s^*_j\right) + K_i \cdot |s'_i - s_i| + \frac{|s'_i - s_i|}{\epsilon} \cdot M^U_i =$$

$$\frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot \pi_i\left({s^*}_i, s^*_j\right) + K_i \cdot |s'_i - s_i| + \frac{|s'_i - s_i|}{\epsilon} \cdot \left(\pi_i\left({s^*}_i, s^*_j\right) - D_i\right) =$$

$$\pi_i\left({s^*}_i, s^*_j\right) + \frac{\epsilon - |s'_i - s_i|}{\epsilon} \cdot K_i \cdot |s'_i - s_i| - \frac{|s'_i - s_i|}{\epsilon} \cdot D_i \leq \pi_i\left({s^*}_i, s^*_j\right) + K_i \cdot |s'_i - s_i| - \frac{|s'_i - s_i|}{\epsilon} \cdot D_i =$$

$$\pi_i\left({s^*}_i, s^*_j\right) + |s'_i - s_i| \cdot \left(K_i - \frac{D_i}{\epsilon}\right) < \pi_i\left({s^*}_i, s^*_j\right),$$

where the first inequality is due to the convexity of $\pi_i(s_i, s_j)$ with respect to s_j, the second inequality is due to $\pi_i\left(s'_i, s^p_j\right) \leq M^U_i$, the third inequality is due to the Lipschitz continuity, the penultimate inequality is implied by $\frac{\epsilon - |s'_i - s_i|}{\epsilon} < 1$, and the last inequality is due to defining ϵ to satisfy $\epsilon < \min\left(\frac{D_i}{K_i}, \frac{D_j}{K_j}\right)$. This proves that player i cannot gain from his deviation, and that $\left(\left(\psi^\epsilon_1, \psi^\epsilon_2\right), (s_1, s_2)\right)$ is a strong BBE.

References

Basu, K. (1994): "The traveler's dilemma: Paradoxes of rationality in game theory," *American Economic Review*, 84(2), 391–395.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 119(3), 861–898.

Costa-Gomes, M., V. P. Crawford, and B. Broseta (2001): "Cognition and behavior in normal-form games: An experimental study," *Econometrica*, 69(5), 1193–1235.

Dekel, E., J. C. Ely, and O. Yilankaya (2007): "Evolution of preferences," *Review of Economic Studies*, 74(3), 685–704.

Heller, Y., and D. Sturrock (2017): "Commitments and partnerships," mimeo.

Holmstrom, B. (1982): "Moral hazard in teams," *The Bell Journal of Economics*, 13(2), 324–340.

Maynard-Smith, J., and G. Price (1973): "The logic of animal conflict," *Nature*, 246, 15–18.

Milgrom, P., and J. Roberts (1990): "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica*, 58(6), 1255–1277.

Nagel, R. (1995): "Unraveling in guessing games: An experimental study," *American Economic Review*, 85(5), 1313–1326.

Stahl, D. O., and P. W. Wilson (1994): "Experimental evidence on players' models of other players," *Journal of Economic Behavior and Organization*, 25(3), 309–327.