Biased-Belief Equilibrium*

Yuval Heller[†] Eyal Winter[‡]

February 21, 2019

Abstract

We investigate how distorted, yet structured, beliefs can persist in strategic situations. Specifically, we study two-player games in which each player is endowed with a biased-belief function that represents the discrepancy between a player's beliefs about the opponent's strategy and the actual strategy. Our equilibrium condition requires that: (1) each player choose a best-response strategy to his distorted belief about the opponent's strategy, and (2) the distortion functions form best responses to one another. We obtain sharp predictions and novel insights into the set of stable outcomes and their supporting stable biases in various classes of games.

JEL classification: C73, D83.

Keywords: commitment, indirect evolutionary approach, distortions, wishful thinking, strategic complements, strategic substitutes.

Standard models of equilibrium behavior attribute rationality to players at two different levels: beliefs and actions (see, e.g., Aumann and Brandenburger, 1995). Players are assumed to behave as if they form correct beliefs about the opponents' behavior, and they choose actions that maximize their

*The authors are very grateful to the anonymous referees for very helpful comments and suggestions.

[†]Department of Economics, Bar Ilan University, Israel. yuval.heller@biu.ac.il. URL: https://sites.google.com/site/yuval26/. The author is grateful to the *European Research Council* for its financial support (ERC starting grant #677057).

[‡]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/. The author is grateful to the German-Israeli Foundation for Scientific Research and Google for their financial support.

utility given the beliefs that they hold. Much of the literature in behavioral and experimental economics that documents violations of the assumption that players have correct beliefs ascribes these violations to cognitive limitations. However, in interactive environments where one person's beliefs affect other persons' actions, belief distortions are not arbitrary, and they may arise to serve some strategic purpose.

In this paper we investigate how distorted, yet structured, beliefs can persist in strategic situations. Our basic assumption here is that distorted beliefs can persist because they offer a strategic advantage to those who hold them even when these beliefs are wrong. More specifically, players often hold distorted beliefs as a form of commitment device that affects the behavior of their counterparts. The precise cognitive process that is responsible for the formation of beliefs is complex, and it is beyond the scope of this paper to outline it. We believe, however, that in addition to analytic assessment of evidence, preferences in the form of desires, fears, and other emotions contribute to the process and, to an extent, facilitate belief biases. If the evidence is unambiguous and decisive, or if the consequence of belief distortion is detrimental to the player's welfare, preferences may play less of a role and learning may work to calibrate beliefs to reality. But when beliefs are biased in ways that favor their holders by affecting the behavior of their counterparts, learning can actually reinforce biases rather than diminish them.

Biased Beliefs Standard equilibrium notions in game theory draw a clear line between preferences and beliefs. The former are exogenous and fixed; the latter can be amended through Bayesian updating but are not allowed to be affected by preferences. However, phenomena such as wishful thinking (see, e.g., Babad and Katz, 1991) and overconfidence (see, e.g., Forbes, 2005; Barber and Odean, 2001; Malmendier and Tate, 2005; Heller, 2014), where beliefs are tilted toward what their holder desires reality to be, suggest that in real life, beliefs and preferences can intermingle, and that biased beliefs may be persistent. Similarly, belief rigidity and belief polarization (see, e.g., Lord, Ross, and Lepper, 1979; Ross and Anderson, 1982) refer to situations in which two people with conflicting prior beliefs each strengthen their beliefs in response to observing the same data. The parties' aversion to depart from their original beliefs can also be regarded as a form of interaction between preferences and beliefs.

It is easy to see how the belief biases described above can have strategic benefits in interactive situations. Wishful thinking and optimism can facilitate cooperation in interactions that require mutual trust. Overconfidence can deter competitors, and belief rigidity can allow an agent to support a credible threat. An important objective of our analysis is to identify the strategic environments that support biases such as wishful thinking as part of equilibrium behavior. It is worthwhile to note that individuals are not the only ones susceptible to strategically motivated belief biases. Governments are prone to be affected by such biases as well. The Bush administration's unsubstantiated confidence in Saddam Hussein's possession of "weapons of mass destruction" prior to the Second Gulf War and the vast discrepancy between Israeli and US intelligence assessments of Iran's nuclear intentions prior to the signing of the Iran nuclear deal can be interpreted as strategically motivated belief distortion.[1]

Belief biases in strategic environments are also connected to self-interest biases regarding moral and ethical standards. Babcock and Loewenstein (1997) had participants in a lab experiment negotiate a deal between a plaintiff and a defendant in a court case. When they asked participants to make predictions about the outcome of the real court case the authors found a significant belief divergence depending on the role participants were assigned to in the negotiations. A similar moral hypocrisy was revealed by Rustichini and Villeval (2014) who showed that subjects' subjective judgments regarding fairness in bargaining depended on the bargaining power they were assigned in the experiment.

A different body of empirical evidence consistent with strategic beliefs

[1]There are other possible interpretations of these controversial real-life examples. In a dynamic real-life setup it is hard to have access to agents' private information, and therefore it is very difficult to achieve direct empirical evidence for persistent biased beliefs. There are a few lab experiments that elicit subjects' beliefs (using monetary incentives and proper scoring rules) about the expected behavior of the opponent. Nyarko and Schotter (2002) demonstrate that the elicited forecasts of subjects about about the opponents' future behavior substantially differ from the empirical play of opponents in the past. Palfrey and Wang (2009) present evidence that forecasts by players (about the opponent's behavior in a simple two-player game) are significantly different from the forecasts of external observers. Moreover, the players' forecasts are systematically biased, and significantly less accurate than the forecasts of the external observers.

is offered by the psychiatric literature on "depressive realism" (e.g., Dobson and Franche, 1989). This literature compares probabilistic assessments conveyed by psychiatrically healthy people with those suffering from clinical depression. Participants in both categories were requested to assess the likelihood of experiencing negative or positive events in both public and private setups. Comparing subjects' answers with the objective probabilities of these events revealed that in a public setup clinically depressed individuals were more realistic than their healthy counterparts for both types of events. The apparent belief bias among healthy individuals can be reasonably attributed to the strategic component of beliefs. Mood disorders negatively affect strategic reasoning (Inoue, Tonooka, Yamada, and Kanba, 2004), which, to a certain extent, may diminish strategic belief distortion among clinically depressed individuals relative to their healthy counterparts.

For biased beliefs to yield a strategic advantage to the agents holding them, it is essential that (1) agents be committed to follow their biased beliefs, and (2) agents best-reply to the perceived behavior induced by their counterparts' biases (both on and off the equilibrium path). For the sake of tractability, we shall avoid formalizing a concrete dynamic model that describes how biased beliefs are formed, and how agents credibly commit to these biased beliefs. Instead, we shall adopt a static approach by imposing equilibrium conditions on the agents' beliefs and the opponents' interpretation of their beliefs. (We discuss our modeling approach and its evolutionary interpretation in Section 2.6, and we present a formal evolutionary foundation in Appendix B.) This static approach is consistent with a large part of the literature on endogenous preferences (see, e.g., the literature cited below). Nevertheless, we mention a few mechanisms that can facilitate these processes and turn biased beliefs into a credible commitment device.

1. Refraining from accessing or using biased sources of information, e.g., subscribing to a newspaper with a specific political orientation, consulting biased experts, and reading Facebook's personalized news feeds, which are typically biased due to friends who hold similar beliefs.

2. Passionately following a religion, a moral principle, or an ideology that has belief implications on human behavior.

3. Possessing personality traits that have implications on beliefs (e.g., narcissism or naivety).

The mechanisms described above are likely not only to induce belief biases, but also to generate signals sent to the player's counterparts about these biases with a certain degree of verifiability. These mechanisms, the signals they induce, and their interpretation are the main forces that facilitate biased-belief equilibrium.

Solution Concept Our notion of biased-belief equilibrium (henceforth, BBE) uses a two-stage paradigm. In the first stage each player is endowed with a biased-belief function. This function represents the discrepancy between a player's beliefs about the strategy profile of other players and the actual profile. In the second stage the players play the biased game induced by their distortion functions, in which each player chooses a best-reply strategy to his biased belief about the opponent's strategy (the chosen strategy profile is referred to as the equilibrium outcome). Finally, our equilibrium condition requires that the distortion functions not be arbitrary, but form best replies to one another.

If one of the players deviates to being endowed with a different biased-belief function, then there might be multiple Nash equilibria in the new biased game induced by this deviation. Our weak notion (*weak BBE*) requires the deviator to be outperformed in at least one equilibrium of the new biased game. Our strong notion (*strong BBE*) requires (1) each agent to have a *monotone* biased belief, according to which he assigns a higher probability to his opponent playing a certain strategy than this probability actually is, and (2) a deviator to be outperformed in *all* Nash equilibria of the new biased game. Our main notion, *BBE*, lies in between these two notions, and it requires (1) each agent to have a monotone biased belief, and (2) the deviator to be outperformed in at least one plausible Nash equilibrium of the new biased game, where we rule out implausible Nash equilibria in which the non-deviator behaves differently even though he does not observe any change in the deviator's perceived strategy.

In Section 2.5 we present our main evolutionary interpretation of our solution concept, according to which the endowed biased beliefs are the result of an evolutionary process of social learning (the interpretation is formalized in Appendix B). In addition, we present an alternative, delegation interpretation of the model (which is formalized in Appendix C).

Nash Equilibrium and BBE We begin our analysis by studying the relations between BBE outcomes and Nash equilibria. We show that any Nash equilibrium can be implemented as the outcome of a BBE, though in some cases this requires that the players have biased beliefs that are accurate on the equilibrium path, but that they be blind to some deviations of the opponent off the equilibrium path. This, in particular, implies that every game admits a BBE. Next, we show that introducing biased beliefs does not change the set of equilibrium outcomes in games in which at least one of the players has a dominant action. By contrast, BBE admits non-Nash behavior in most other games.

Main Results Our main results show that the notion of BBE induces substantial predictive power in various classes of interval games. In these classes of games the strategy of each player is a number in a bounded interval, where a higher strategy (interpreted as a higher investment) induces a higher payoff for the opponent. We begin by characterizing the set of BBE in games with strategic complements (Bulow, Geanakoplos, and Klemperer, 1985), such as price competition with differentiated goods (Example 2), input games (Example 9 in Appendix A.4), and stag hunt games (Example 10 in Appendix A.4). We show three key properties of any BBE: (1) *overinvestment*: the strategy of each agent is (weakly) higher than the best reply to the opponent's (real) strategy, (2) *ruling out bad outcomes*: both players invest more than their investments in the worst Nash equilibrium of the underlying game, and (3) *wishful thinking*: each agent perceives his opponent as investing (weakly) more than the opponent's real investment.

Next, we characterize the set of BBE in games with strategic substitutes, such as Cournot competitions (Example 3) and hawk-dove games (Example 5 in Appendix A.5). We show three key properties of any BBE: (1) *underinvestment*: the strategy of each agent is (weakly) higher than the best reply to the opponent's (real) strategy, (2) *ruling out excellent outcomes*: at least one of the players invests less than his investments in one of the Nash equilibria of the underlying game, and (3) *wishful thinking*: each agent perceives his opponent as investing (weakly) more than the opponent's real investment.

Finally, we characterize the set of BBE in a class of games (which are less common in economic interactions), in which the strategy of player 1 is a complement of player 2's strategy, while the strategy of player 2 is a substitute of player 1's strategy (e.g., duopolistic competition in which one firm chooses its quantity while the opposing firm chooses its price (Singh and Vives, 1984), and various classes of asymmetric contests (Dixit, 1987)). We show that in this class of games agents present *pessimism* in any BBE: each agent perceives his opponent as investing (weakly) less than the opponent's real investment.

Additional Results Our next result shows an interesting class of BBE that exist in all games. We say that a strategy is undominated Stackelberg if it maximizes a player's payoff in a setup in which the player can commit to an undominated strategy, and his opponent reacts by best-replying to this strategy. We show that every game admits a BBE in which one of the players is "strategically stubborn" in the sense of having a constant belief about the opponent's strategy, and always playing his undominated Stackelberg strategy, while the opponent is "rational" in the sense of having undistorted beliefs and best-replying to the player's true strategy.

Section 6.2 shows that unless one imposes both requirements on the definition of a BBE, namely, monotonicity and ruling out implausible equilibria, then the set of BBE outcomes is very large in various classes of games. Specifically, Proposition 8 shows that for a large class of finite games, a strategy profile is a monotone weak BBE iff (1) no player uses a strictly dominated strategy, and (2) the payoff of each player is above the minmax payoff of the player in a setup in which both players are restricted to choose only undominated strategies (i.e., strategies that are not strictly dominated). Proposition 9 shows a similar folk theorem result for non-monotone strong BBE in a large class of interval games.

Empirical Predictions Our main results imply two empirical predictions. First, they suggest that efficient (non-Nash equilibrium) outcomes are easier to support in games with strategic complements, relative to games with strategic substitutes. This prediction is consistent with the experimental findings of Potters and Suetens (2009), which show that there is significantly more cooperation in games with strategic complements than in the case of strategic substitutes.

Our second empirical prediction is that wishful thinking is strategically stable in many common environments, though some (less common) strategic interactions may induce pessimism. This empirical prediction is consistent with the experimental evidence that people tend to present wishful thinking, while the presented level of wishful thinking may substantially differ between various environments; see, e.g., Babad and Katz (1991); Budescu and Bruderman (1995); Bar-Hillel and Budescu (1995) and Mayraz (2013).

Structure The structure of this paper is as follows. We discuss the related literature in Section 1. Section 2 describes the model. In Section 3 we analyze the relations between BBE and Nash equilibria. Section 4 defines games with strategic complements/substitutes and wishful thinking. We analyze these games and present our main results in Section 5. In Section 6 we present additional results: (1) the relation between BBE and strategies played by a Stackelberg leader, and (2) folk theorem results when relaxing the definition of BBE. We conclude in Section 7. All the appendices of the paper appear in the online supplementary material. 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.

1 Related Literature and Contributions

Our paper aims at making a contribution to the behavioral game theory literature. Much of this literature concerns behavioral equilibrium concepts that depart from the framework of Nash equilibrium by introducing weaker rationality conditions. This has been done primarily at the level of preferences (e.g., Güth and Yaari, 1992; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Acemoglu and Yildiz, 2001; Heifetz, Segev, et al., 2004; Dekel, Ely, and Yilankaya, 2007; Heifetz, Shannon, and Spiegel, 2007a; Friedman and Singh, 2009; Herold and Kuzmics, 2009; Heller and Winter, 2016; Winter, Garcia-Jurado, and Mendez-Naya, 2017). But it has also been done at the level of beliefs (e.g., Geanakoplos, Pearce, and Stacchetti, 1989; Rabin, 1993; Battigalli and Dufwenberg, 2007; Attanasi and Nagel, 2008; Battigalli and Dufwenberg, 2009; Battigalli, Dufwenberg, and Smith, 2015; Gannon and Zhang, 2017). This latter literature deals with belief-dependent preferences, and focuses primarily on the way players' beliefs about the intentions of others affect their preferences and behavior.

Our equilibrium concept also operates on beliefs rather than preferences but is based on an inherently different approach. Preferences in our model are not affected by beliefs but beliefs are biased in a way that serves players' strategic purposes. Our analysis of biased belief goes beyond characterizing equilibrium outcomes. An additional important objective is to identify the belief biases that support these equilibrium outcomes in different strategic environments. Central to our analysis are belief-distortion properties, such as wishful thinking and pessimism, that sustain BBE in different strategic environments.

The existing literature has presented various prominent solution concepts that assume that players have distorted beliefs. Some examples include models of level-k and cognitive hierarchy (see, e.g., Stahl and Wilson, 1994; Nagel, 1995; Costa-Gomes, Crawford, and Broseta, 2001; Camerer, Ho, and Chong, 2004), analogy-based expectation equilibrium (Jehiel, 2005), cursed equilibrium (Eyster and Rabin, 2005), and Berk-Nash equilibrium (Esponda and Pouzo, 2016). These equilibrium notions have been helpful in understanding strategic behavior in various setups, and yet these notions pose a conceptual challenge to our understanding of the persistence of distorted beliefs, even in view of the empirical evidence for such persistence. If players can infer the truth ex post why don't they calibrate their beliefs toward reality? Much of the literature presenting such models points to cognitive limitations as the source of this rigidity. Our model and analysis offer an additional perspective to this issue by suggesting that belief biases that yield a strategic advantage in the long run are likely to emerge in equilibrium. In this sense our approach can be viewed as providing a tool to explain why some cognitive limitations persist while others do not (see Example 11 in Appendix A, in which we show how level-1 behavior can be supported as part of a BBE outcome in the traveler's dilemma).

Our notion of BBE is related to the notion of conjectural equilibrium (Battigalli and Guaitoli, 1997, originally written in 1988) insofar as both solution concepts relax the Nash equilibrium's requirement that beliefs need to be consistent with actual play (while still requiring that an agent's action has to be optimal given the agent's belief). A conjectural equilibrium is defined in an environment in which players do not observe each other's actions but rather observe signals of each other's actions, according to an exogenous feedback correspondence. In a conjectural equilibrium each player best replies to his belief about the opponent's action, and this belief is required to be consistent with the signal observed by the player. There are two key structural differences between a BBE and a conjectural equilibrium. First, a BBE is defined in an environment in which there is no exogenous feedback correspondence; rather, the feedback correspondence is implicitly defined as part of the solution concept by the agents' biased-belief functions. These biased-belief functions are not restricted by a consistency requirement with respect to an exogenous feedback mechanism, but rather they are are restricted by the requirement that each biased-belief function has to be a best reply against the opponent's biased belief. The second structural difference is that while a BBE describes what would be the agent's belief for *any* feasible action of the opponent, a conjectural equilibrium describes only the agent's belief about the equilibrium action of the opponent.

Despite these structural differences, it is interesting to discuss relations between the equilibrium behavior induced by each solution concept, i.e., the relations between a BBE outcome and a conjectural equilibrium outcome. Without restricting the feedback correspondence, the notion of conjectural equilibrium is rather broad (it rules out only strictly dominated strategies), and, accordingly, any BBE outcome is a conjectural equilibrium outcome. Fudenberg and Levine's (1993) notion of self-confirming equilibrium deals with extensive-form games, and refines conjectural equilibrium by requiring that the feedback correspondence is the one in which each player observes the opponent's realized actions (but does not observe the opponent's behavior off the equilibrium path). In the setup of two-player one-shot games, which is the focus of the present paper, the set of self-confirming equilibria coincides with the set of Nash equilibria (whereas the set of BBE outcomes is broader and includes non-Nash outcomes). Another refinement of conjectural equilibrium is the rationalizable conjectural equilibrium (Rubinstein and Wolinsky, 1994; the notion has been generalized to games with structural uncertainty in Esponda, 2013). This concept requires that the agents' beliefs be consistent with the common knowledge that all players maximize utility given their signals. There is no inclusion relation between the set of BBE outcomes and the set of rationalizable conjectural equilibrium outcomes. Specifically, in games with a unique rationalizable action profile, such as price competitions with differentiated goods and Cournot competitions, the unique rationalizable conjectural equilibrium outcome is the Nash equilibrium (for any feedback correspondence), while the set of BBE outcomes is substantially larger (see Examples 2 and 3). By contrast, in games such as stag hunt and hawk–dove, when the feedback correspondence is non-informative any action profile is a conjectural equilibrium outcome, while the set of BBE outcomes is much more restricted (see Examples 10 and 12 in Appendix A).

2 Model

2.1 Underlying Game

Let $i \in \{1,2\}$ be an index used to refer to one of the players in a two-player game, and let j be an index referring to the opponent. Let $G = (S, \pi)$ be a normal-form two-player game (henceforth, *game*), where $S = (S_1, S_2)$ and each S_i is a convex compact set of strategies. Specifically, we focus on two cases:

1. *Finite games*: Each S_i is a simplex over a finite set of pure actions, where each strategy corresponds to a mixed action (i.e., A_i is a finite set of pure actions, and $S_i = \Delta(A_i)$), and the von Neumann–Morgenstern payoff function is linear with respect to the mixing probability.

2. *Interval games*: Each S_i is a bounded interval in $\mathbb{R}$ (e.g., each player chooses a real number representing quantity, price, or effort).

We denote by $\pi = (\pi_1, \pi_2)$ players' payoff functions; i.e., $\pi_i : S \to \mathbb{R}$ is a function assigning each player a payoff for each strategy profile. We use s_i to refer to a typical strategy of player i. We assume each payoff function $\pi_i(s_i, s_j)$ to be continuously twice differentiable in both parameters and weakly concave in the first parameter (s_i).

Let BR (resp., BR^{-1}) denote the (inverse) best-reply correspondence; i.e.,

$$BR(s_i) = argmax_{s_j \in S_j}(\pi_j(s_i, s_j))$$

is the set of best replies against strategy $s_i \in S_i$, and

$$BR^{-1}(s_i) = \{s_j \in S_j | s_i \in BR(s_j)\}$$

is the set of strategies for which s_i is a best reply against them.

In a finite game, we use $a_i \in A_i$ to denote also the degenerate mixed action that assigns mass one to a_i. When the set of actions of a player is given as an ordered set $A_i = (a_i^1, a_i^2, ..., a_i^n)$, we identify a mixed action with a vector $s_i = (\alpha_1, \alpha_2, ..., \alpha_n)$, where $0 \le \alpha_k = s_i\left(a_i^k\right)$ for each $1 \le k \le n$, and $\sum_k \alpha_k = 1$. Given two strategies $s_i, s_i' \in S_j$ and $\alpha \in [0,1]$, let $\alpha \cdot s_i + (1-\alpha) \cdot s_i'$ be the mixture of the two strategies: $(\alpha \cdot s_i + (1-\alpha) \cdot s_i')(a_i) = \alpha \cdot s_i(a_i) + (1-\alpha) \cdot s_i'(a_i)$.

When there are two (ordered) actions for each player (say, $A_i = \{c_i, d_i\}$), we identify a mixed action s_i with the probability it assigns to the first pure action $s_i(c_i)$, and we identify the set of strategies S_i with the interval $[0,1]$. Thus, a game with two actions for each player can be captured both as a finite game and as an interval game.

2.2 Biased-Belief Function

We start here with the definition of biased-belief functions that describe how players' beliefs are distorted. A *biased belief* $\psi_i : S_j \rightarrow S_j$ is a *continuous* function that assigns to each strategy of the opponent, a (possibly distorted) belief about the opponent's play. That is, if the opponent plays s_j, then player i believes that the opponent plays $\psi_i(s_j)$. We call s_j the opponent's real strategy, and we call $\psi_i(s_j)$ the opponent's perceived (or biased) strategy. Formally, the continuity requirement is that if $(s_{j,n})_n \underset{n\to\infty}{\rightarrow} s_j$, then $(\psi_i(s_{j,n}))_n \underset{n\to\infty}{\rightarrow} \psi_i(s_j)$ (where in a finite game, we say that $(s_{j,n})_n \underset{n\to\infty}{\rightarrow} s_j$ iff $(s_{j,n}(a))_n \underset{n\to\infty}{\rightarrow} s_j(a)$ for each action a).

Remark 1. Two reasons motivate us to require that a biased belief be continuous: (1) continuity implies that each biased game (defined below) admits a Nash equilibrium, which allows us to simplify the definition of BBE, and (2) continuity reflects a plausible restriction that a small change in the opponent's strategy should induce a small change in the perceived strategy. In Appendix D we present an alternative (and somewhat more complicated) definition of a BBE that relaxes the assumption that biased beliefs must be continuous, and we show that all the BBE characterized in the results of the paper remain BBE when we allow deviators to use discontinuous biased beliefs.

We say that a biased belief $\psi_i : S_j \rightarrow S_j$ is *monotone* if:

1. In interval games: $s_j \geq s'_j$ implies $\psi_i(s_j) \geq \psi_i\left(s'_j\right)$ for each strategy $s_j \in S_j$.

2. In finite games: If the opponent plays a_j more often, while keeping the same proportion of playing the remaining actions, then the perceived probability that the opponent plays any other action weakly decreases (which implies, in particular, that the perceived probability that the opponent plays a_j weakly increases); that is,

$$(\psi_i((1-\alpha)\cdot s_j + \alpha \cdot a_j))\left(a'_j\right) \leq (\psi_i(s_j))\left(a'_j\right)$$

   for each $\alpha \in [0,1]$, each action $a_j \in A_j$, each action $a'_j \neq a_j$, and each strategy $s_j \in \Delta(A_j)$. In particular, when the game has two actions for each player, a biased belief ψ_i is monotone iff ψ_i is weakly increasing in α_j; i.e., $\alpha_j \geq \alpha'_j$ implies that $\psi_i(\alpha_j) \geq \psi_i\left(\alpha'_j\right)$.

Monotone biased beliefs reflect a plausible restriction on the distortion of agents, namely, that if the opponent changes his real strategy in some direction, the agent captures the direction of the change correctly, but may have the wrong perception about the magnitude of the change.

Let I_d be the undistorted (identity) function, i.e., $I_d(s) = s$ for each strategy s. A biased belief ψ is *blind* if the perceived opponent's strategy is independent of the opponent's real strategy, i.e., if $\psi(s_j) = \psi\left(s'_j\right)$ for each $s_j, s'_j \in S_j$. With a slight abuse of notation we use s_i to denote also the blind biased belief ψ_j that is always equal to s_i.

2.3 Biased Game

An underlying game and a profile of biased beliefs jointly induce a biased game in which the (biased) payoff of each player is determined by the perceived strategy of the opponent. Formally:

Definition 1. Given an underlying game $G = (S, \pi)$ and a profile of biased beliefs (ψ_i, ψ_j), let the *biased game* $G_\psi = (S, \psi \circ \pi)$ be defined as the game with the following payoff function $(\psi \circ \pi)_i : S_i \times S_j \to \mathbb{R}$ for each player i:

$$(\psi \circ \pi)_i (s_i, s_j) = \pi_i (s_i, \psi_i (s_j)).$$

A Nash equilibrium of a biased game is defined in the standard way. Formally, a pair of strategies $s^* = (s_1^*, s_2^*)$ is a Nash equilibrium of a biased game $G_\psi = (S, \psi \circ \pi)$, if each s_i^* is a best reply against the perceived strategy of the opponent, i.e.,

$$s_i^* = argmax_{s_i \in S_i} \left(\pi_i \left(s_i, \psi_i \left(s_j^*\right)\right)\right).$$

Let $NE(G_\psi) \subseteq S_1 \times S_2$ denote the set of all Nash equilibria of the biased game G_ψ.

Observe that the set of strategies of a biased game is convex and compact, and the payoff function $(\psi \circ \pi)_i : S_i \times S_j \to \mathbb{R}$ 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 biased game G_ψ admits a Nash equilibrium (i.e., $NE\left(G_{\left(\psi_i',\psi_j^*\right)}\right) \neq \emptyset$.)

2.4 Weak and Strong BBE

We are now ready to define our equilibrium concept. A weak biased-belief equilibrium (abbr. weak BBE) is a pair consisting of a profile of biased beliefs and a profile of strategies, such that: (1) each strategy is a best reply to the perceived strategy of the opponent, and (2) each biased belief is a best reply to the opponent's biased belief, in the sense that any agent who chooses a different biased-belief function is outperformed in at least one equilibrium in the new biased game (relative to the agent's payoff in the original equilibrium). Formally:

Definition 2. A *weak BBE* is a pair (ψ^*, s^*), where $\psi^* = (\psi_1^*, \psi_2^*)$ is a profile of biased beliefs and $s^* = (s_1^*, s_2^*)$ is a profile of strategies satisfying: (1) $\left(s_i^*, s_j^*\right) \in NE\left(G_{\psi^*}\right)$, and (2) for each player i and each biased belief ψ_i', there exists a strategy profile $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i',\psi_j^*\right)}\right)$, such that the following inequality holds: $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

The notion of weak BBE is arguably too permissive because it allows incumbents: (1) to have implausible non-monotone beliefs, and (2) to outperform the deviators in a single Nash equilibrium of the biased game (while, possibly, the incumbents are outperformed by the deviators in many other equilibria). Proposition 8 (in Section 6.2.2) demonstrates that this single Nash equilibrium, in which the deviators are outperformed, may be implausible due to allowing the incumbents to "discriminate" against the deviators, even though the deviators exhibit exactly the same perceived behavior as the rest of the population.

The more restrictive refinement of strong BBE requires that (1) incumbents have monotone beliefs, and (2) deviators who choose a different biased-belief function be outperformed in *all* equilibria of the induced biased game. Formally:

Definition 3. A *weak BBE* (ψ^*, s^*) is a *strong BBE* if (1) each biased function ψ_i^* is monotone, and (2) the inequality $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$

holds for every player i, every biased belief ψ_i', and every strategy profile $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$.

2.5 BBE

Finite games typically induce multiple Nash equilibria. This is often the case also with respect to biased games. This suggests that the refinement of strong BBE may be too restrictive, as there are are potentially many Nash equilibria of many biased games, and the requirement of the deviators being outperformed in all these equilibria might be too demanding. Our main solution concept, BBE, lies in between weak BBE and strong BBE.

In a BBE, the deviator is required to be outperformed in at least one *plausible* equilibrium of the new biased game. Roughly speaking, in a plausible equilibrium of the new biased game induced by a deviation of player i to a different biased belief, player j is allowed to choose a new strategy only if he distinguishes between i's original strategy and i's new strategy. More precisely, implausible equilibria are defined as follows. We say that a Nash 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) player j playing his original strategy induces 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 4. Given weak BBE (ψ^*, s^*), deviating player i, and biased belief ψ_i', we say that a Nash equilibrium of the biased game $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ is *implausible* if: (1) $\psi_j^*\left(s_i'\right) = \psi_j^*\left({s^*}_i\right)$, (2) $s_j^* \neq s_j'$, and (3) $\left(s_i', s_j^*\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$. An equilibrium is *plausible* if it is not implausible. Let $PNE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ be the set of all plausible equilibria of the biased game $G_{\left(\psi_i', \psi_j^*\right)}$.

Note that it is immediate from Definition 4 and the nonemptiness of $NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ that $PNE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$ is nonempty.

Definition 5. Weak BBE (ψ^*, s^*) is *a BBE* if (1) each biased function ψ_i^* is monotone, and (2) for each player i and each biased belief ψ_i', there exists a plausible Nash equilibrium $\left(s_i', s_j'\right) \in PNE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$, such that $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$.

A strategy profile $s^* = (s_1^*, s_2^*)$ is a *(resp., strong, weak) BBE outcome* if there exists a profile of biased beliefs $\psi^* = (\psi_1^*, \psi_2^*)$ such that (ψ^*, s^*) is a (resp., strong, weak) BBE. In this case we say that the biased belief ψ^* supports (or implements) the outcome s^*.

2.6 Discussion of the Model

Evolutionary/Learning Interpretation Biases can emerge in a learning process that reinforces biases that yield a strategic advantage to their holders. Specifically, we interpret a BBE to be a reduced-form solution concept capturing the essential features of an evolutionary process of cultural or social learning. Our methodology follows the extensive literature that studies the stability of endogenous preferences using the "indirect evolutionary approach" (see, e.g., Güth and Yaari, 1992; Güth, 1995; Fershtman and Weiss, 1998; Dufwenberg and Güth, 1999; Koçkesen, Ok, and Sethi, 2000; Guttman, 2003; Güth and Napel, 2006; Heifetz, Shannon, and Spiegel, 2007b; Friedman and Singh, 2009; Herold and Kuzmics, 2009; Alger and Weibull, 2013; Heller and Mohlin, 2017). We apply this modeling approach to the study of endogenous biased beliefs in a setup in which biased beliefs induce behavior, behavior determines "success," and success regulates the evolution of biased beliefs.

In Appendix B we formally adapt the definition of a stable population state from Dekel, Ely, and Yilankaya (2007) to our setup, and show that the adapted definition is equivalent to a strong BBE. In what follows we briefly and informally present our evolutionary interpretation. Consider two large populations of agents: agents who play the role of player 1, and agents who play the role of player 2. In each round agents from each population are randomly matched to play a two-person game against opponents from the other population. Each agent in each population is endowed with a biased-belief function. For simplicity, we focus on "homogeneous" populations, in which all agents in the population have the same monotone biased-belief function. Agents distort their perception about the behavior of the agents in the other population according to their endowed biased-belief functions, and they play a Nash equilibrium of the biased game.

With small probability a few agents ("mutants") in one of the populations (say, population 1) may be endowed with a different biased-belief function due to a random error or experimentation. We assume that agents of population 2 observe whether their opponents are mutants or not, and that the agents of population 2 and the mutants of population 1 gradually adapt their play against each other into an equilibrium of the new biased game. Note that a dynamic adaptation into playing a Nash equilibrium of the biased game requires agents of population 2 to know the perceived strategy currently being played by the mutants of population 1, but the agents do not need to know the biased beliefs of the mutants of population 1.

Finally, we assume that the total "success" (fitness) of agents is monotonically influenced by their (unbiased) payoff in the underlying game, and that there is a slow process in which the composition of the population evolves. This slow process might be the result of a slow flow of new agents who join the population. Each new agent randomly chooses one of the incumbents in his own population as a "mentor" (and mimics the mentor's biased belief), where the probabilities are such that agents with higher fitness are more likely to be chosen as mentors. If the original population state is not a BBE, it implies that there are mutants who outperform the remaining incumbents in their own population, which in turn implies that the original population state is not stable, as new agents are likely to mimic more successful mutants. By contrast, if the original population state is a BBE, it implies that for any mutant there is a new equilibrium in which the mutants are weakly outperformed relative to the incumbents of their own population, and this can allow the BBE to remain a stable state (as illustrated in the detailed example in Appendix B.3).

Variants of the Solution Concept *The main solution concept we use in the paper is BBE.* In Section 6.2 we demonstrate that unless one applies both requirements of Definition 5, namely, monotonicity and ruling out implausible equilibria, then the set of BBE is very large (folk theorem results), and some of the biased beliefs that support some of these equilibria seem implausible. The intuition for the monotonicity requirement is quite straightforward (ruling out peculiar biased beliefs in which an opponent who deviates to play a higher strategy is perceived as deviating to play a lower strategy). The second requirement rules out implausible equilibria in which a player responds to his opponent's deviation in spite of not being able to perceive it

In what follows we sketch a dynamic justification for the second requirement of ruling out implausible equilibria (following the evolutionary interpretation described above). Consider a BBE $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$. Assume that both (s_1', s_2') and (s_1', s_2^*) are Nash equilibria of the biased game $G_{(\psi_1', \psi_2^*)}$. In what follows, we briefly, and informally, explain why (s_1', s_2') is not a plausible equilibrium of the new biased game (and, thus, why it is ruled out in the definition of BBE). Consider a deviation of some agents in the population playing in the role of player 1 to having the biased belief ψ_1'. Following this deviation, strategy s_1^* might not be a best reply to the perceived strategy of player 2 (i.e., $s_1^* \notin BR(\psi_1'(s_2^*))$) and, as a result, the deviating agents might change their strategy to s_1', which is a best reply to the perceived strategy of player 2 (i.e., $s_1' \in BR(\psi_1'(s_2^*))$). The current strategy profile (s_1', s_2^*) is a Nash equilibrium of the biased game (i.e., $(s_1', s_2^*) \in NE\left(G_{(\psi_1', \psi_2^*)}\right)$). In order to move from this equilibrium to (s_1', s_2'), agents of population 2, who are matched against the deviators, have to change their behavior from s_2^* to s_2', but there is no reason for them to do so, as their current behavior (namely, s_2^*) is already a best reply to the perceived strategy of the deviators (i.e., $s_2^* \in BR(\psi_1^*(s_1'))$), as well as being how they are used to playing against non-deviators.

Delegation Interpretation A different interpretation of our solution concept relies on strategic delegation. The literature on strategic delegation (see, e.g., Fershtman, Judd, and Kalai, 1991; Dufwenberg and Güth, 1999; Fershtman and Gneezy, 2001) deals with players who strategically use other agents to play on their behalf, where the agents so used may have different preferences than the players using them. We adapt this approach to our setup in which agents differ in their biased beliefs (rather than in their preferences). Specifically, in Appendix C we show that the notion of weak

BBE is equivalent to a subgame-perfect equilibrium of a two-stage game in which in stage one each unbiased player strategically chooses the biased belief of his agent, and in the second stage the biased agents play on behalf of the players (and each agent can observe the opposing agent's biased beliefs).

Partial Observability The requirement that an agent be able to observe that his opponent belongs to a group of "mutant" agents who have different biased beliefs than the rest of the population can be explained by pre-play social cues and messages that facilitate this observation. In Appendix E we show that this observability need not be perfect. We generalize the model to partial observability by studying a setup in which, when an agent is matched with a mutant opponent, the agent privately observes the opponent to be a mutant with probability $0 < p \leq 1$. We show that all our results hold in this extended setup for p sufficiently close to one (and some of the results hold also for low levels of p).

3 Nash Equilibria and BBE Outcomes

In this section we study the relations between Nash equilibria and BBE outcomes.

3.1 Nash Equilibria and Biased Beliefs

We begin with a simple observation that shows that in any weak BBE in which the outcome is not a Nash equilibrium, at least one of the players must distort the opponent's perceived strategy. The reason for this observation is that if both players have undistorted beliefs, then it must be that each agent best-replies to the opponent's strategy, which implies that the outcome is a Nash equilibrium of the underlying game.

The following example demonstrates that even Nash equilibria may require distorted beliefs to be supported as BBE outcomes. Specifically, Example 1 shows that this is the case for Nash equilibrium in a Cournot competition. The intuition behind Example 1 is straightforward. The Cournot equilibrium cannot be supported by undistorted beliefs because such pairs of beliefs will induce one of the players to adopt a distorted belief by which he expects his opponent not to produce at all, and to best-reply to this distorted belief by producing the monopoly quantity. This in turn will force the opponent to reduce his production substantially below the Cournot level, making the deviator better off.

Example 1 (*Cournot equilibrium cannot be supported by undistorted beliefs, yet it can be supported by blind beliefs*)**.** Consider the following symmetric Cournot game $G = (S, \pi)$: $S_i = [0, 1]$ and $\pi_i (s_i, s_j) = s_i \cdot (1 - s_i - s_j)$ for each player i. The interpretation of the game is as follows. Each s_i is interpreted as the quantity chosen by firm i, the price of both goods is determined by the linear inverse demand function $p = 1 - s_i - s_j$, and the marginal cost of each firm is normalized to be zero. The unique Nash equilibrium of the game is $s_i^* = s_j^* = \frac{1}{3}$, which yields a payoff of $\frac{1}{9}$ to both players. Assume to the contrary that this outcome can be supported as a weak BBE by the undistorted beliefs $\psi_i^* = \psi_j^* = I_d$. Consider a deviation of player 1 to the blind belief $\psi_1' \equiv 0$. The unique equilibrium of the biased game $G_{(0,I_d)}$ is $s_1' = \frac{1}{2}$, $s_2' = \frac{1}{4}$, which yields a payoff of $\frac{1}{8} > \frac{1}{9}$ to the deviator. The unique Nash equilibrium $s_i^* = s_j^* = \frac{1}{3}$ can be supported as the outcome of the strong BBE $\left(\left(\frac{1}{3}, \frac{1}{3}\right), \left(\frac{1}{3}, \frac{1}{3}\right)\right)$ with blind beliefs, in which each agent believes the opponent is playing $\frac{1}{3}$ regardless of the opponent's actual play, and the agent plays the unique best reply to this belief, which is the strategy $\frac{1}{3}$.

Remark 2 (Interpretation of Nash equilibria supported by blind beliefs.). We interpret an undistorted belief as describing an agent who has an accurate belief about the opponent's behavior on the equilibrium path, and, in addition, the agent keeps looking for cues that his opponent might have a different type, and if the agent observes such a cue, the agent evaluates the opponent's likely behavior, and best-replies to this assessment. Example 1 shows that the Cournot equilibrium cannot be supported by a population in which each agent keeps looking for cues for his opponent's type. In such a population, deviators would strictly earn by having a blind biased belief that induces the deviator to play the Stackelberg strategy. The incumbents will identify the mutants' type, and they will respond by playing the Stackelberg follower action, which will benefit the deviators.

By contrast, the second part of Example 1 (and its generalization in Proposition 1 below) shows that any Nash equilibrium can be supported by a blind belief, which is accurate on the equilibrium path. We interpret such a belief as describing an agent who understands correctly the equilibrium behavior of the opposing player, and ignores signals that suggest that his opponent is about to do something else. Our observation that it is rather equilibrium that supports belief rigidity, a prevalent behavioral phenomenon, and not disequilibrium is, we believe, quite interesting.

3.2 Any Nash Equilibrium is a BBE Outcome

The following result generalizes the second part of Example 1, and shows that any (strict) Nash equilibrium is an outcome of a (strong) BBE in which both players have blind beliefs that are accurate on the equilibrium path.

Proposition 1. *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 $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a (strong) BBE.*

Proof. The fact that (s_1^*, s_2^*) is a Nash equilibrium of the underlying game implies that (s_1^*, s_2^*) is an equilibrium of the biased game $G_{(\psi_1^*, \psi_2^*)}$. The fact that the beliefs are blind implies that for any biased belief ψ_i', there is an equilibrium in the biased game $G_{(\psi_i', \psi_j^*)}$ in which player j plays s_j^* and player i gains at most $\pi_i\left(s_i^*, s_j^*\right)$, which implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE. Moreover, if (s_1^*, s_2^*) is a strict equilibrium, then in any equilibrium of any biased game $G_{(\psi_i', \psi_j^*)}$, player j plays s_j^* and player i gains at most $\pi_i\left(s_i^*, s_j^*\right)$, which implies that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a strong BBE. □

An immediate corollary of Proposition 1 is that every game admits a BBE.

Corollary 1. *Every game admits a BBE.*

3.3 Zero-Sum Games

Recall that a game is *zero sum* if there exists $c \in R^+$ such that $\pi_i(s_i, s_j) + \pi_j(s_i, s_j) = c$ for each strategy profile $(s_i, s_j) \in S$.

The following simple result shows that the unique Nash equilibrium payoff of a zero-sum game is also the unique payoff in any weak BBE.

Claim 1. The unique Nash equilibrium payoff of a zero-sum game is also the unique payoff in any weak BBE.

Proof. Let v_i be the unique Nash equilibrium payoff of player i in the underlying zero-sum game. Assume to the contrary that there exists a weak BBE (ψ^*, s^*) in which the payoff of player i is strictly lower than v_i. Consider a deviation of player i into the undistorted bias function $\psi_i' = I_d$. The assumption that (ψ^*, s^*) is a weak BBE implies that the deviator gets strictly less than v_i in a Nash equilibrium $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$, but this is impossible as the definition of v_i implies that there exists $\hat{s}_i$ satisfying $\pi_i\left(\hat{s}_i, s_j'\right) \geq v_i > \pi_i\left(s_i', s_j'\right)$. □

Example 6 in Appendix A.2 shows that even though 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.

3.4 Games with a Dominant Strategy

Next we show that if at least one of the players has a dominant strategy, then any weak BBE outcome must be a Nash equilibrium. Formally:

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

Proof. Observe that s_i^* is the unique best reply of player i to any perceived strategy of player j, and, as a result, player i plays the dominant action s_i^* in any weak BBE. Assume to the contrary that there is a weak BBE in which player j does not best-reply against s_i^*. Consider a deviation of player j to choosing the undistorted belief I_d. Observe that player i still plays his dominant action s_i^*, and that player j best-replies to s_i^* in any Nash equilibrium of the induced biased game, and, as a result, player j achieves a strictly higher payoff, and we get a contradiction. □

Proposition 2 implies, in particular, that defection is the unique weak BBE outcome in the prisoner's dilemma game. Example 7 in Appendix

A.1 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, allows us to sustain cooperation as a strong BBE outcome.

4 Monotone Games and Wishful Thinking

In this section we present a large class of games with monotone externalities and monotone differences, and define the notions of wishful thinking and pessimism, which will be analyzed in Section 5.

4.1 Monotone Games

We say that an interval game is monotone if it satisfies two conditions:

1. *Monotone externalities*: the payoff function of each player is strictly monotone in the opponent's strategy. Without loss of generality, we assume that the *externalities* are *positive*, i.e., the payoff of each player is increasing in the opponent's strategy, i.e., that $\frac{\partial \pi_i(s_i,s_j)}{\partial s_j} > 0$ for each player i and each pair of strategies s_i, s_j. The assumption of positive externalities (given monotone externalities) is indeed without loss of generality because if originally the externalities with respect to player j are negative, then we can redefine player j's strategy to be its inverse, and obtain positive externalities; for example, defining the difference between maximal capacity and quantity to be the strategy of each player in a Cournot competition yields a game with positive externalities.
   In a game with positive externalities we refer to a player's strategy as his *investment*, and when s_i increases we refer to this increase a larger investment by as player i.

2. *Monotone differences*: For each player i, the derivative of the player's payoff with respect to his own strategy (i.e., $\frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$) is strictly monotone in the opponent's strategy. Specifically, we divide the set of monotone games into three disjoint and exhaustive subsets:

(a) *Strategic complements (increasing differences, supermodular games)*: $\frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$ is strictly increasing in s_j for each player i and each strategy s_i (or, equivalently, $\frac{\partial^2 \pi_i(s_i,s_j)}{\partial s_i \cdot \partial s_j} > 0$ for each s_i, s_j). Games with strategic complements are common in the economics literature, and include, in particular, price competitions with differentiated goods (Example 2), input games (Example 9 in Appendix A.4), and stag-hunt games (Example 10 in Appendix A.4). Finite games with a payoff structure that resembles a discrete variant of strategic complements include the traveler's dilemma (Example 11 in Appendix A.4).

(b) *Strategic substitutes (decreasing differences, submodular games)*: $\frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$ is strictly decreasing in s_j for each player i and each strategy s_i (or, equivalently, $\frac{\partial^2 \pi_i(s_i,s_j)}{\partial s_i \cdot \partial s_j} < 0$ for each s_i, s_j). Games with strategic *substitutes* are common in the economics literature, and include, in particular, Cournot (quantity) competitions (Example 3 below) and hawk-dove games (see Example 12 in Appendix A.5).

(c) *Opposing differences*: $\frac{\partial \pi_i(s_i,s_j)}{\partial s_i}$ is decreasing in s_j (for each strategy s_i), while $\frac{\partial \pi_j(s_i,s_j)}{\partial s_j}$ is increasing in s_i (for each strategy s_j). Games with opposing differences are less common in the economics literature. Examples of these games include (1) duopolies in which one firm chooses its quantity, while the other firm chooses its price (see, e.g., Singh and Vives, 1984), and (2) asymmetric contests, in which it is often the case that a commitment of the favorite (underdog) player to exert more (less) effort induces the opponent to exert less effort (see, e.g., Dixit, 1987).

4.2 Wishful Thinking

We say that player i exhibits wishful thinking if the perceived opponent's strategy yields a higher payoff to the player relative to the real strategy the opponent plays. Formally:

Definition 6. Player i *exhibits wishful thinking* in weak BBE $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ if $\pi_i\left(s_i, \psi_i^*\left(s_j\right)\right) \geq \pi_i\left(s_i, s_j^*\right)$ for each $s_i \in S_i$.

Remark 3. Note that in a game with positive externalities player i exhibits wishful thinking in weak BBE $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ iff $\psi_2^*(s_1^*) \geq s_1^*$ and $\psi_1^*(s_2^*) \geq s_2^*$.

Similarly, we define the opposite notion, that of exhibiting pessimism. We say that a BBE exhibits pessimism if the perceived opponent's strategy yields a lower payoff to the player relative to the real opponent's strategy for all strategy profiles. It exhibits pessimism in equilibrium if it satisfies this property with respect to the strategy the opponent plays on the equilibrium path. Formally:

Definition 7. A weak BBE $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ *exhibits pessimism* if $\pi_i(s_i, \psi_i^*(s_j)) \leq \pi_i\left(s_i, s_j^*\right)$ for all $s_i \in S_i$.

4.3 Additional Definitions

In what follows we present two definitions that will be used in the analysis in the following sections: undominated Pareto optimality, and biased-belief minmax payoff.

We say that a strategy profile is undominated Pareto optimal if it is (1) undominated, and (2) Pareto optimal among all undominated strategy profiles. Formally:

Definition 8. Strategy profile (s_1^*, s_2^*) is *undominated Pareto optimal if (1) $s_i^* \in S_i^U$ for each player i, and (2) there does not exist $(s_1', s_2') \in S_1^U \times S_2^U$ with a payoff that Pareto dominates (s_1^*, s_2^*), i.e., $\pi_1(s_1^*, s_2^*) \leq \pi_1(s_1', s_2')$ and $\pi_2(s_1^*, s_2^*) \leq \pi_2(s_1', s_2')$ where at least one of these inequalities is strict.*

A biased-belief minmax payoff for player i (denoted by $\tilde{M}_i^U$) is the maximal payoff player i can guarantee to himself in the following process: (1) player j chooses an arbitrary perceived strategy of player i, and (2) player i chooses a strategy profile, under the constraint that player j's strategy is a best reply to the perceived strategy chosen above. That is, $\tilde{M}_i^U$ is the payoff player i can guarantee himself no matter how his opponent (player j, she) perceives player i's action, assuming that player j best-replies to what he believes player i is doing (and if there are multiple best replies, then we assume that player j chooses the best reply that is optimal for player i). Formally:

Definition 9. Given game $G = (A, u)$, let $\tilde{M}_i^U$, the *biased-belief minmax payoff* of player i, be defined as follows:

$$\tilde{M}_i^U = \min_{s_i' \in S_i^U} \left(\max_{(s_i, s_j) \in S_i \times BR\left(s_i'\right)} \pi_i \left(s_i, s_j\right) \right).$$

Observe that the biased-belief minmax is weakly larger than the undominated maxmin (Definition 10), i.e., $\tilde{M}_i^U \geq M_i^U$ with an equality if the strategy of player j that guarantees that player i's payoff is at most M_i^U is a unique best reply against some strategy of player i (which is the case, in particular, if the payoff function is strictly concave).

5 Main Results

Our main results characterize the set of BBE and BBE outcomes in three classes of games: games with strategic complements, games with strategic substitutes, and games with strategic opposites.

5.1 Preliminary Result: Necessary Conditions for a Weak BBE Outcome

We begin by defining undominated strategies and the undominated minmax payoff, which will be used to characterize necessary conditions for a strategy profile to be a weak BBE outcome.

Strategy s_i of player i is *undominated* if it is a best reply of some strategy of the opponent, i.e., if there exists strategy $s_j \in S_j$, such that $s_i \in BR(s_j)$. We say that a strategy profile is *undominated* if both strategies in the profile are undominated. Recall that in a finite game, due to the minmax theorem, a strategy is undominated iff it is not strictly dominated by another strategy.

Let $S_i^U \in S_i$ denote the *set of undominated strategies* of player i. Observe that S_i^U is not necessarily a convex set.

An undominated minmax payoff for player i is the maximal payoff player i can guarantee to himself in the following process: (1) player j chooses an arbitrary undominated strategy, and (2) player i chooses a strategy (after observing player j's strategy). Formally:

Definition 10. Given game $G = (S, u)$, let M_i^U , the *undominated minmax payoff* of player i, be defined as follows:

$$M_i^U = \min_{s_j \in S_j^U} \left(\max_{s_i \in S_i} \pi_i \left(s_i, s_j \right) \right).$$

Observe that the undominated minmax is weakly larger than the standard maxmin, i.e., $M_i^U \geq \min_{s_j \in S_j} \left(\max_{s_i \in S_i} \pi_i \left(s_i, s_j \right) \right)$ with an equality if player j does not have any strictly dominated strategy[2] (i.e., if $S_j^U = S_j$).

The following simple result (which will be helpful in deriving the main results in the following subsections) shows that any weak BBE outcome is an undominated strategy profile that yields a payoff above the player's undominated minmax payoff to each player.

Proposition 3. *If a strategy profile $s^* = (s_1^*, s_2^*)$ is a weak BBE outcome, then (1) the profile s^* is undominated and (2) $\pi_i (s^*) \geq M_i^U$.*

Proof. Assume that $s^* = (s_1^*, s_2^*)$ is a biased-belief equilibrium outcome. This implies that each s_i^* is a best reply to the player's distorted belief, which implies that each s_i^* is undominated. Assume to the contrary, that , $\pi_i (s^*) < M_i^U$. Then, by deviating to the undistorted function I_d, player i can guarantee a fitness of at least M_i^U in any distorted equilibrium. □

5.2 Games with Strategic Complements

Our first main result characterizes the set of BBE outcomes in games with strategic complements. It shows that a strategy profile is a BBE outcome essentially iff (I) it is undominated, (II) it yields a payoff above the undominated/biased-belief minmax payoff to both players, and (III) both players overinvest (i.e., use a weakly higher strategy than the best reply to the opponent). Formally:

[2] The undominated minmax payoff might be strictly higher than the undominated *maxmin* payoff due to the non-convexity of S_U^j; i.e., player i might be able to guarantee only a lower payoff in a setup in which player j is allowed to choose his undominated strategy after observing player i's chosen strategy.

Proposition 4. *Let G be a game with* strategic complements and positive externalities.

1. *Let (s_1^*, s_2^*) be a BBE outcome. Then (s_1^*, s_2^*) has the following properties: (I) it is undominated, and it satisfies for each player i: (II) $\pi_i\left(s_i^*, s_j^*\right) \geq M_i^U$, and (III) overinvestment: $s_i^* \geq \min\left(BR\left(s_j^*\right)\right)$.*

2. *Let (s_1^*, s_2^*) be an undominated profile that satisfies, 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, (s_1^*, s_2^*) is a BBE outcome.*
   Moreover, if $\pi_i(s_i, s_j)$ is strictly concave in s_i (i.e., $\frac{\partial \pi_i^2(s_i,s_j)}{\partial s_i^2} > 0$) then (s_1^*, s_2^*) is a strong BBE outcome.

Sketch of Proof (formal proof in Appendix F.1).
Part 1: Proposition 3 implies (I) and (II). To prove (III, overinvestment), assume to the contrary that $s_i^* < \min\left(BR\left(s_j^*\right)\right)$. Consider a deviation of player i that induces him to invest slightly more than s_i^*. The fact that $s_i^* < \min\left(BR\left(s_j^*\right)\right)$ implies that player i strictly earns from his own deviation. The assumption that the biased belief of the opponent is monotone implies that the agent's deviation induces the opponent to invest more and, thereby to further improve the agent's payoff. Thus, the agent gains from the deviation, and (s_1^*, s_2^*) cannot be a BBE outcome.

Part 2: The strategy profile (s_1^*, s_2^*) is supported as a BBE outcome by a profile of biased beliefs (ψ_1^*, ψ_2^*) in which each biased belief ψ_j^* satisfies: (1) blindness to good news: ψ_j^* distorts any $s_i' \geq s_i^*$ into $BR^{-1}\left(s_j^*\right)$, and (2) overreaction to bad news: ψ_j^* distorts any $s_i' < s_i^*$ to a sufficiently low strategy $\psi_j(s_i')$, such that player i loses in any strategy profile $\left(s_i', s_j'\right)$ in which player j best-replies to the perceived strategy of player i (i.e., $s_j' \in BR(\psi_j(s_i'))$). These properties imply that $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ is a BBE (and a strong BBE if the payoff function is strictly concave). □

Recall that a game with strategic complements admits a *lowest Nash equilibrium* $(\underline{s}_1, \underline{s}_2)$ in which both players invest less than in any other Nash equilibrium, i.e., $s_i' \geq \underline{s}_i$ for each player i and each strategy s_i' that is played in a Nash equilibrium (see, e.g., Milgrom and Roberts, 1990).

An immediate corollary of Prop. 4 is that in each BBE outcome, both players invest more than in any Nash equilibrium. Formally:

Corollary 2. *Let G be a game with strategic complements and positive externalities with a lowest Nash equilibrium $(\underline{s}_1, \underline{s}_2)$ that satisfies $\underline{s}_1 <$* $\max(S_i)$ *for each player i. Let (s_1^*, s_2^*) be a BBE outcome. Then $\underline{s}_i \leq s_i^*$ for each player i.*

Proof. The result is immediate from part (1.III) of Proposition 4 (namely, that both agents weakly overinvest in any BBE outcome), and the observation (which is formally proved in Lemma 1 in Appendix F.2) that $s_i^* < \underline{s}_i$ implies that at least one of the players strictly underinvests. □

Corollary 2 shows that the notion of BBE rules out socially bad outcomes in which one (or both) of the players invests less effort than the lowest Nash equilibrium. In particular, in a price competition with differentiated goods (see Example 2 below), the corollary implies that the price chosen by any player in any BBE is at least the player's price in the unique Nash equilibrium of the game.

The final corollary shows the close relation between BBE and wishful thinking. Specifically, it shows that any biased belief in any BBE (with a non-extreme outcome) of a game with strategic complements exhibits wishful thinking. The intuition is that wishful thinking causes an agent to believe that the opponent is playing a higher action, which induces the agent to respond with a higher action, which, in turn, causes the opponent to respond by playing a higher action, which benefits the agent.[3]

Corollary 3. *Let G be a game with positive externalities and strategic complements. Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a BBE. If $s_i^* \notin \{\min(S_i), \max(S_i)\}$,* then player i exhibits wishful thinking (i.e., $\psi_i^*\left(s_j^*\right) \geq s_j^*$).

Proof. Assume to the contrary that $\psi_i^*\left(s_j^*\right) < s_j^*$. The strategic complementarity implies that $\max\left(BR\left(\psi_i^*\left(s_j^*\right)\right)\right) \leq \min\left(BR\left(s_j^*\right)\right)$ with an

[3]Corollary 3 allows for pessimism of player i in a BBE only if player i plays an extreme strategy (either, the minimal feasible strategy or the maximal feasible strategy) and his pessimism does not affect his play; i.e., the best reply against the real opponent's strategy and the best reply against the perceived opponent's strategy coincide in being the same extreme strategy. For example, this is the case in the biased beliefs that support the action profile (s_i, s_j) in the stag hunt game analyzed below.

equality only if

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

(see Lemma 2 in Appendix F.3 for a formal proof of this claim). Part 1 of Proposition 4 and the definition of a BBE imply that

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

The previous inequalities jointly imply that

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

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

Next, we apply our analysis of games with strategic complements to price competition with differentiated goods (the linear city model à la Hotelling). Specifically, we show that (1) players choose prices above the unique Nash equilibrium price in all BBE, and (2) any undominated symmetric price profile above the Nash equilibrium price can be supported as a strong BBE. In Appendix A.4 we present three additional examples: input games, stag hunt games, and the traveler's dilemma.

Example 2 (*Price competition with differentiated goods; see a textbook analysis in Mas-Colell, Whinston, and Green, 1995, Section 12.C*)**.** Consider a mass one of consumers equally distributed in the interval $[0, 1]$. Consider two firms that produce widgets, located at the two extreme locations: 0 and 1. Every consumer wants at most one widget. Producing a widget has a constant marginal cost, which we normalize to be zero. Each firm i chooses price $s_i \in [0, M]$ for its widgets. The total cost of buying a widget from firm i is equal to its price s_i plus t times the consumer's distance from the firm, where $t \in [0, M]$). Each buyer buys a widget from the firm with the lower total buying cost. This implies that the total demand for good i is given by function $q_i(s_i, s_j)$:

$$q_i(s_i, s_j) = \begin{cases} 0 & \frac{s_j - s_i + t}{2\cdot t} < 0 \\ \frac{s_j - s_i + t}{2\cdot t} & 0 < \frac{s_j - s_i + t}{2\cdot t} < 1 \\ 1 & \frac{s_j - s_i + t}{2\cdot t} > 1, \end{cases}$$

The payoff (profit) of firm i is given by $\pi_i(s_i, s_j) = s_i \cdot q_i(s_i, s_j)$. Observe that the payoff function is strictly concave in s_i for any non-extreme s_j (and it is weakly concave for the extreme values of s_j). One can show that the game has strategic complements, and that the best-reply function of each player is:

$$s_i(s_j) = \begin{cases} \frac{s_j + t}{2} & s_j < 3 \cdot t \\ s_j - t & s_j \geq 3 \cdot t. \end{cases}$$

It is well known that the unique Nash equilibrium of this example is given by $s_i = s_j = t$, which yields a payoff of $\frac{t}{2}$ to each firm.

Observe that the set of undominated strategies of each player is the interval $\left[\frac{t}{2}, \frac{M+t}{2}\right]$ (where $\frac{t}{2}$ is the best reply against 0 and $\frac{M+t}{2}$ is the best reply against M). This implies that the undominated minmax of each player is equal to $\pi_i\left(\frac{3}{4} \cdot t, \frac{t}{2}\right) = \frac{3}{4} \cdot t \cdot \frac{3}{8} = \frac{9}{32} \cdot t$. Proposition 4 implies that a strategy profile (s_i, s_j) is a BBE outcome if for each player i: (1) $s_i \in \left[\frac{t}{2}, \frac{M+t}{2}\right]$ (undominated strategy), (2) $\pi_i(s_i, s_j) > \frac{9}{32} \cdot t$ (payoff above the undominated minmax payoff),[4] and (3) overinvestment: $s_i \geq \frac{s_j + t}{2}$.

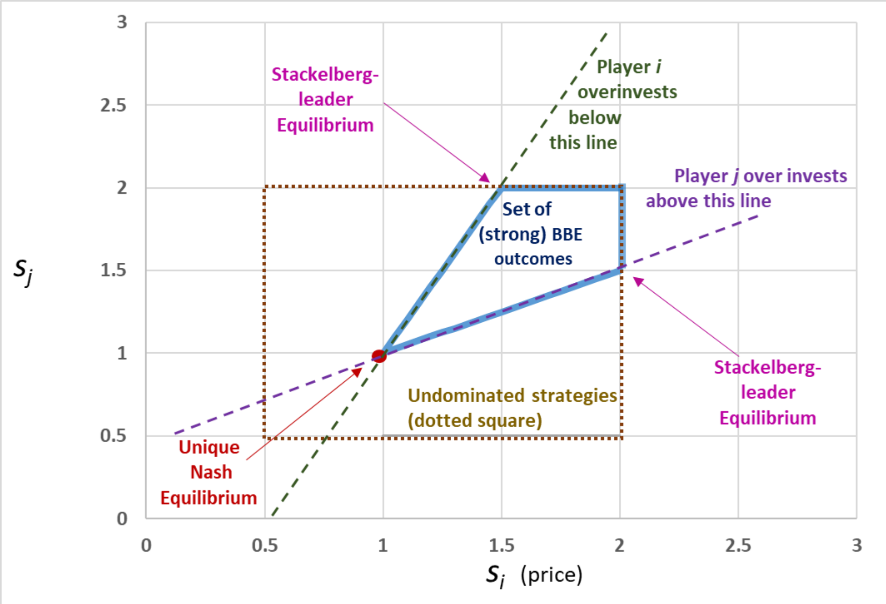
Figure 1 shows the set of BBE outcomes (which coincides with the set of strong BBE outcomes, due to the strict concavity of the payoff function), for $t = 1$ and $M = 3$.

Observe that the sum of the payoffs to the two firms, $s_i \cdot q_i(s_i, s_j) + s_j \cdot q_j(s_i, s_j)$, is a mixed average of s_i and s_j. The fact that the Nash equilibrium is in the bottom left corner of the set of BBE outcomes implies that all BBE outcomes (except the Nash equilibrium itself) strictly improve

[4] One can show that the constraint on s_i implied by $\pi_i(s_i, s_j) > \frac{9}{32} \cdot t$ is nonbinding. The constraint is

$$s_i \in \left(\frac{s_j + t - \sqrt{(s_j + t)^2 - 2.25 \cdot t^2}}{2}, \frac{s_j + t - \sqrt{(s_j + t)^2 - 2.25 \cdot t^2}}{2} \right).$$

Figure 1: The Set of (Strong) BBE Outcomes in Example 2 ($t = 1$, $M = 3$)



social welfare relative to the Nash equilibrium (as measured by the sum of payoffs of the two firms).

Next, we make two observations regarding the implications of the extent of wishful thinking on the players' payoffs (both observations hold also for the input games in Example 9 in Appendix A.4):

1. Increasing the wishful thinking of both players improves the players' payoffs. Specifically, with respect to symmetric BBE outcomes, a higher level of wishful thinking induces a higher equilibrium price and a higher payoff to the players: a wishful thinking level of $x^* \equiv \psi^*(s^*) - s^* \in [0,1]$ induces the symmetric BBE price $x^* + t = x^* + 1$ (which is implied by the perceived bast-reply equation $s^* = \frac{\psi^*(s^*)+t}{2} = \frac{s^*+x^*+t}{2}$), which yields a payoff of $\frac{x^*+1}{2}$ to each player.

2. When the wishful thinking levels of the two players differ, the player with the higher wishful thinking level has a lower payoff. This is because the difference between the payoffs of a firm with price s_i and an opponent with price $s_j < s_i$ is equal to:

$$\pi_i - \pi_j = s_i \cdot \left(\frac{s_j - s_i + 1}{2}\right) - s_j \cdot \left(\frac{s_i - s_j + 1}{2}\right) = 0.5\left(s_j\left(s_j - 1\right) - s_i\left(s_i - 1\right)\right) < 0.$$

Intuitively, wishful thinking is like a public good in this setup: (1) a higher level of wishful thinking is beneficial to social welfare, and (2) if the two players have different levels of wishful thinking, the player with the higher level obtains a lower payoff.

We conclude the example by presenting a symmetric biased belief $\psi_1^* = \psi_2^*$ that supports the outcome $(2,2)$ as the BBE $((\psi_1^*, \psi_2^*), (2,2))$ in the game with $M = 3$ and $t = 1$:

$$\psi_i^*(s_j) = \begin{cases} 3 & s_j > 2 \\ 2 \cdot s_j - 1 & s_j \in [0.5, 2] \\ 0 & s_j < 0.5. \end{cases}$$

Observe that: (1) this BBE yields a payoff of 1 to each player and (2) the biased belief presents wishful thinking, i.e., $\psi_i^*(2) = 3 > 2$. Further observe that a player with biased belief ψ_i^* plays the same strategy as the opponent (regardless of the opponent's biased belief) in any equilibrium of the biased game in which the opponent plays any intermediate value of s_j (i.e., $s_j \in [0.5, 2]$) :

$$s_i(\psi_i^*(s_j)) = \begin{cases} s_i(3) = 2 & s_j > 2 \\ s_i(2 \cdot s_j - 1) = 0.5 \cdot (2 \cdot s_j - 1 + 1) = s_j & s_j \in [0.5, 2] \\ s_i(0) = 0.5 & s_j < 0.5. \end{cases}$$

This implies that the equilibrium payoff of a deviating player j who plays strategy s_j is equal to:

$$\pi_j(s_j, s_i(\psi_i^*(s_j))) = \begin{cases} s_j \cdot 0.5 \cdot (2 - s_j + 1) = s_j \cdot 0.5 \cdot (3 - s_j) < 1 & s_j > 2 \\ s_j \cdot 0.5 \cdot q(s_j, s_j) = 0.5 \cdot s_j & s_j \in [0.5, 2] \\ s_j \cdot 0.5 \cdot (0.5 - s_j + 1) = s_j \cdot 0.5 \cdot (1.5 - s_j) < 0.25 & s_j < 0.5, \end{cases}$$

and it is at most 1, which implies that a deviator cannot gain from his deviation.

Finally, note that Figure 1 shows that the two Stackelberg-leader equilibria (the unique subgame-perfect equilibrium of the sequential games in which one of the players plays first, and the opponent replies after observing the leader's strategy) are included in the set of BBE, as is proven in general in Proposition 7.

5.3 Games with Strategic Substitutes

Our next result characterizes the set of BBE outcomes in games with strategic substitutes (and positive externalities). It shows that a strategy profile is a BBE outcome essentially iff (I) it is undominated, (II) it yields a payoff above the undominated/biased-belief minmax payoff to both players, and (III) both players underinvest (i.e., use a weakly lower strategy than the best reply to the opponent). Formally:

Proposition 5. *Let G be a game with* strategic substitutes and positive externalities*.*

1. *Let (s_1^*, s_2^*) be a BBE outcome. Then (s_1^*, s_2^*) has the following properties: (I) it is undominated, and if satisfies 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 that satisfies, 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, (s_1^*, s_2^*) is a BBE outcome.*
   Moreover, if $\pi_i(s_i, s_j)$ is strictly concave then (s_1^*, s_2^*) is a strong BBE outcome.

The proof, which is analogous to the proof of Proposition 4, is presented in Appendix F.4.

An immediate corollary of Proposition 4 is that in each BBE outcome, at least one of the players invests less relative to his maximal Nash equilibrium investment. Formally:

Corollary 4. *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 $s_i^e \geq s_i^*$.*

Proof. The result is immediate from part (1-III) of Proposition 4 (namely, that both agents weakly underinvest in any BBE outcome), and the observation (which is formally proved in Lemma 3 in Appendix F.5) that if the effort of each player s_i^* is strictly below all of his Nash equilibrium efforts, then at least one of the players strictly underinvests. □

Corollary 4 shows that the notion of BBE rules out socially good outcomes in which both players invest more effort relative to their maximal Nash equilibrium effort. In particular, in a Cournot competition (see Example 3 below), the corollary implies that a collusive outcome in which both players retain more unused capacity relative to the unique Nash equilibrium.

Combining Corollary 2 and Corollary 4 implies the following *empirical prediction of our model and the notion of BBE: efficient (non-Nash equilibrium) outcomes are easier to support in games with strategic complements, relative to games with strategic substitutes.* This prediction is consistent with the experimental findings of Potters and Suetens (2009), which show that there is significantly more cooperation in games with strategic complements than in games with strategic substitutes.

The following corollary shows that in games with strategic substitutes, as in games with strategic complements, there is the a close relation between BBE and wishful thinking. Specifically, it shows that any biased belief in any BBE (with a non-extreme outcome) of a game with strategic substitutes exhibits wishful thinking. The intuition is that wishful thinking causes an agent to believe that the opponent is playing a higher action, which induces the agent to respond with a lower action, which, in turn, causes the opponent to respond by playing a higher action, which benefits the agent.

Corollary 5. *Let G be a game with positive externalities and strategic substitutes. Let $((\psi_1^*, \psi_2^*), (s_1^*, s_2^*))$ be a BBE. If $s_i^* \notin \{\min(S_i), \max(S_i)\}$,* then player i exhibits wishful thinking (i.e., $\psi_i^*\left(s_j^*\right) \geq s_j^*$).

The proof, which is analogous to the proof of Corollary 3, is presented in Appendix F.6.

The following example characterizes the set of BBE outcomes in a Cournot competition. Appendix A.5 presents an analysis of another game of strategic substitutes: the hawk-dove game.

Example 3 (*Cournot competition with linear demand*)**.** Consider a symmetric Cournot competition, where we relabel the set of strategies to describe unused capacity, rather than quantity, in order to follow the normalization of positive externalities. Formally, let $G = (S, \pi)$: $S_i = [0,1]$ and $\pi_i(s_i, s_j) = (1 - s_i) \cdot (s_i + s_j - 1)$ for each player i. Each s_i is interpreted as the unused capacity (= one minus the quantity, i.e., $s_i = 1 - q_i$) chosen by firm i, the price of both goods is determined by the linear inverse demand function $p = 1 - q_i - q_j = s_i + s_j - 1$, and the marginal cost of each firm is normalized to be zero.

Observe that:

1. $BR(s_i) = 1 - \frac{s_i}{2}$, and the unique Nash equilibrium of the game is $s_1^* = s_2^* = \frac{2}{3}$, which yields a payoff of $\frac{1}{9}$ to both players.

2. The set of undominated strategies of each player is the interval $[0.5, 1]$ (where 1 is the best reply against 0, and 0.5 is the best reply against 1).

3. The symmetric Pareto optimal profile (which is also undominated) is $s_i = s_j = \frac{3}{4}$, yielding a payoff of $\frac{1}{8}$ to each player.

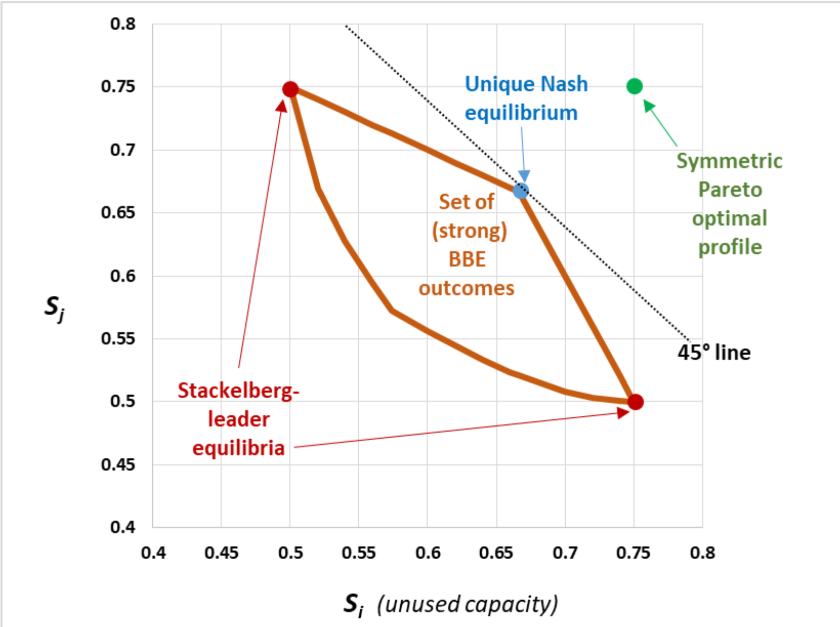
4. The undominated minmax payoff $M_i^U = \frac{1}{16}$, which is achieved by the opponent playing his lowest undominated strategy $s_i = 0.5$.

5. The sum of payoffs of both players when they play profile (s_1, s_2) is $\pi_1(s_1, s_2) + \pi_2(s_1, s_2) = (2 - (s_i + s_j)) \cdot ((s_i + s_j) - 1)$, which is an increasing function of $s_i + s_j$ in the domain of undominated strategies $s_i, s_j \geq 0.5$.

Applying the analysis of the previous subsection to a Cournot competition shows that strategy profile (s_1, s_2) is a BBE outcome iff it satisfies for each player i: (1) the strategy is undominated: $s_i \geq 0.5$, (2) the payoff is greater than the undominated minmax payoff: $(1 - s_i) \cdot (s_i + s_j - 1) \geq \frac{1}{16} = M_i^U$, and (3) underinvestment relative to the best reply against the opponent: $s_i \leq BR(s_j) = 1 - \frac{s_j}{2}$. Due to having a strictly concave payoff function, the set of BBE outcomes coincides with the set of strong BBE outcomes. Figure 1 shows this set of BBE outcomes (the strategy profiles that satisfy the above three conditions).

Observe that the unique Nash equilibrium $\left(\frac{2}{3}, \frac{2}{3}\right)$ is the profile that maximizes the sum $s_i + s_j$ within the set of BBE. This implies that all

Figure 2: The Set of (Strong) BBE Outcomes in a Cournot Competition



other BBE outcomes yield lower social welfare (as measured by the sum of payoffs) relative to the Nash equilibrium.

Next, we make two observations regarding the implications of the level of wishful thinking on the players' payoffs.

1. Increasing the wishful thinking of both players decreases the players' payoffs. Specifically, when focusing on symmetric BBE outcomes, a higher level of wishful thinking induces a lower level of unused capacity and a lower payoff to both players; the higher level of production is induced by the false assessment of each firm that the other firm is producing less than it actually does.[5]

2. When the wishful thinking levels of the two players differ, the player with the higher wishful thinking has a higher payoff. This is because the difference between the payoffs of a firm with price s_i and an opponent with price $s_j < s_i$ is equal to

$$\pi_i - \pi_j = s_i \cdot \left(\frac{s_j - s_i + 1}{2}\right) - s_j \cdot \left(\frac{s_i - s_j + 1}{2}\right) = 0.5\left(s_j\left(s_j - 1\right) - s_i\left(s_i - 1\right)\right) < 0.$$

[5] A wishful thinking level of $x^* \equiv \psi^*(s^*) - s^* \in [0, 0.28]$ induces a symmetric BBE unused capacity of $s^* = \frac{2-x^*}{3}$ (which is implied by the perceived best-reply equation $s^* = 1 - \frac{\psi^*(s^*)}{2} = 1 - \frac{s^*+x^*}{2}$).

Thus, a higher level of wishful thinking is beneficial to social welfare, but harms the player with the higher level (relative to the opponent's payoff).

Finally, note that Figure 2 shows that the two Stackelberg-leader equilibria (the unique subgame-perfect equilibria of the sequential games in which one of the players plays first, and the opponent replies after observing the leader's strategy) are included in the set of BBE, as is proven in general in Proposition 7.

5.4 Pessimism in Games with Opposing Differences

The results of the previous two subsections present a strong tendency of BBE to exhibit wishful thinking both in games with strategic complements and in games with strategic substitutes. This raises the question of which class of games induces pessimism. In this section we show that the answer to this question is games with strategic opposites. Recall that these are games in which the strategy of player 1 is a complement of player 2's strategy, while the strategy of player 2 is a substitute of player 1's strategy, e.g., duopolistic competitions in which one firms chooses its quantity while the opposing firm chooses its price (Singh and Vives, 1984) and various classes of asymmetric contests (Dixit, 1987).

Proposition 6 characterizes the set of BBE outcomes in games with strategic opposites (and positive externalities).

It shows that a strategy profile is a BBE outcome essentially iff (I) it is undominated, (II) it yields a payoff above the undominated/biased-belief minmax payoff to both players, and (III) player 1 (for whom player 2's strategy is a complement) underinvests, while player 2 (for whom player 1's strategy is a substitute) overinvests. Formally:

Proposition 6. *Let G be a game with positive externalities and strategic opposites:* $\frac{\partial^2 \pi_1(s_1,s_2)}{\partial s_1 \partial s_2} > 0$ and $\frac{\partial^1 \pi_2(s_1,s_2)}{\partial s_1 \partial s_2} < 0$ for each pair of strategies s_1, s_2.

1. *Let (s_1^*, s_2^*) be a 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) $s_1^* \leq \max\left(BR\left(s_2^*\right)\right)$ and $s_2^* \geq \min\left(BR\left(s_1^*\right)\right)$ (i.e., player 1 underinvests and player 2 overinvests relative to the best reply to the opponent).*

2. *Let (s_1^*, s_2^*) be a profile satisfying the following conditions: (I) (s_1^*, s_2^*)*

is 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,* $\left(s_1^*, s_2^*\right)$ *is a BBE outcome.*

The proof, which is analogous to the proof of Proposition 4, is presented in Appendix F.7.

The following corollary shows that in games with strategic opposites, there is a close relation between BBE and pessimism. Specifically, it shows that any biased belief in any BBE (with a non-extreme outcome) of a game with strategic opposites exhibits pessimism. The intuition is that pessimism causes player 1 to believe that player 2 is playing a lower action, which induces player 1 to respond with a lower action, which, in turn, causes player 2 to respond by playing a higher action, which benefits player 1. Similarly, pessimism causes player 2 to believe that player 1 is playing a lower action, which induces player 2 to respond with a higher action, which, in turn, causes player 1 to respond by playing a higher action, which benefits player 2.

Corollary 6. *Let* $\left(\left(\psi_1^*, \psi_2^*\right), \left(s_1^*, s_2^*\right)\right)$ *be a BBE of a game with* positive externalities *and strategic opposites (i.e.,*$\frac{\partial^2 \pi_1(s_1,s_2)}{\partial s_1 \partial s_2} > 0$ *and* $\frac{\partial^2 \pi_2(s_1,s_2)}{\partial s_1 \partial s_2} < 0$ *for each pair of strategies* s_1, s_2*). If* $s_i^* \notin \left\{\min\left(S_i\right), \max\left(S_i\right)\right\}$*, then player* i *exhibits pessimism (i.e.,* $\psi_i^*\left(s_j^*\right) \leq s_j^*$*).*

The proof, which is analogous to the proof of Corollary 3, is presented in Appendix F.8.

Next, we present an example of a game with strategic opposites, and we characterize the set of BBE in this game.

Example 4 (*Matching pennies with positive externalities*)**.** The game presented in Table 1, a variant of the matching pennies game, is played as follows:

1. Player 1 (player 2) gains 1 utility point from matching (mismatching) his opponent.

2. Each player i induces a gain of 3 utility points to his opponent by choosing heads (action h_i).

The game admits a unique Nash equilibrium $(0.5, 0.5)$ with a payoff of 1.5 to each player. The (undominated) minmax payoff of each player is 1 (obtained when the opponent plays t_j). Observe that the game has positive externalities, that the strategy of player 2 is a strategic complement for player 1, while the strategy of player 1 is a strategic substitute for player 2.

Table 1: Matching Pennies with Positive Externalities

	h_2	t_2
h_1	$4, 2$	$-1, 4$
t_1	$2, 1$	$1, -1$

Applying the analysis of the previous section shows that the game admits 2 classes of BBE:

1. A class in which the players mix while giving a larger weight to playing heads (the action with positive externalities), pessimism, and one-directional blindness. Specifically, each BBE in this class $((\psi_1^*, \psi_2^*), (\beta_1, \beta_2))$ satisfies for each player i: (I) $\beta_i \in [0.5, 1]$ (i.e., both players play heads more frequently than in the unique Nash equilibrium), (II) pessimism: $\psi_i^*(\beta_j) = 0.5 < \beta_j$, and (III) one-sided blindness: $\psi_1^*(\alpha) = 0.5$ for each $\alpha \geq \beta_2$; $\psi_1^*(\alpha) < 0.5$ for each $\alpha < \beta_1$; $\psi_2^*(\alpha) = 0.5$ for each $\alpha \leq \beta_2$; and $\psi_2^*(\alpha) < 0.5$ for each $\alpha > \beta_2$.

2. A class in which player 1 mixes while giving more weight to tails, while player 2 plays heads. Both players exhibit pessimism. Specifically, each BBE in this class $((\psi_1^*, \psi_2^*), (\beta_1, \beta_2))$ satisfies for each player i: (I) $\beta_1 \in [0, 0.5]$ and $\beta_2 = 1$ (i.e., player 1 plays tails more frequently than in the unique Nash equilibrium, while player 2 always plays heads), (II) pessimism for player 1: $\psi_1^*(\beta_2 = 1) = 0.5 < 1$, and $\psi_2^*(\beta_1) = 0.5$ (player 2 is not pessimistic, due to the fact that he chooses the extreme action 1), and (III) $\psi_2^*(\alpha) > 0.5$ for each $\alpha > \beta_1$.

Observe that any profile (β_1, β_2), where $\beta_2 < 0.5$ or ($\beta_1 < 0.5$ and $\beta_2 < 1$), cannot be a BBE outcome:

1. If $\beta_2 < 0.5$ and $\beta_1 = 0$, then player 2's payoff is negative, and less than his undominated minmax payoff of 1.

2. If $\beta_2 < 0.5$ and $\beta_1 > 0$, then player 1 can gain by deviating to $\psi_1' \equiv 0$, as the only possible equilibria of the new biased game are $(0_1, 0_2)$ and $(0_1, \beta_2)$, both of which induce a higher payoff to player 1 relative to (β_1, β_2).

3. If $\beta_1 < 0.5$ and $\beta_2 < 1$, then player 2 can gain by deviating to $\psi_2' \equiv 0$, as the only possible equilibria of the new biased game are $(0_1, 1_2)$ and $(\beta_1, 1_2)$, both of which induce a higher payoff to player 2 relative to (β_1, β_2).

5.5 Empirical Prediction Regarding Wishful Thinking

Arguably, the class of games with strategic opposites (which induces pessimism) is less common in strategic interactions than the classes of games with strategic complements/substitutes (both of which induce wishful thinking). This observation suggests the following empirical predictions of our model: (1) wishful thinking is more common than pessimism, and (2) there are some (less common) strategic interactions that induce pessimism. This empirical prediction is consistent with the experimental evidence that people tend to present wishful thinking, although, the extent of wishful thinking may substantially differ across different environments and may disappear in some environments (see, e.g., Babad and Katz, 1991; Budescu and Bruderman, 1995; Bar-Hillel and Budescu, 1995; Mayraz, 2013).

6 Additional Results

6.1 BBE with Strategic Stubbornness

In this subsection we present an interesting class of BBE that exist in all games. In this class, one of the players is "strategically stubborn" in the sense that he plays his undominated Stackelberg strategy (defined below) and has blind beliefs, while his opponent is "flexible" in the sense of having unbiased beliefs.

A strategy is undominated Stackelberg if it maximizes a player's payoff in a setup in which the player can commit to an undominated strategy, and his opponent reacts by choosing the best reply that maximizes player i's payoff. Formally:

Definition 11. The strategy s_i is an undominated Stackelberg strategy if it satisfies

$$s_i = \text{argmax}_{s_i \in S_i^U} \left(\max_{s_j \in BR(s_i)} \left(\pi_i \left(s_i, s_j \right) \right) \right).$$

Let $\pi_i^{\text{Stac}} = \max_{s_i \in S_i^U} \left(max_{s_j \in BR(s_i)} \left(\pi_i \left(s_i, s_j \right) \right) \right)$ be the undominated Stackelberg payoff. Observe that $\pi_i^{\text{Stac}} \geq \pi_i \left(s_1^*, s_2^* \right)$ for any Nash equilibrium $\left(s_1^*, s_2^* \right) \in NE\left(G \right)$.

Our next result shows that every game admits a BBE in which one of the players: (1) has a blind belief, (2) plays his undominated Stackelberg strategy, and (3) obtains his undominated Stackelberg payoff. The opponent has undistorted beliefs. Moreover, this BBE is strong if the undominated Stackelberg strategy is a unique best reply to some undominated strategy of the opponent.

The intuition behind Proposition 7 is as follows. The "strategically stubborn" player i cannot gain from a deviation, because player i already obtains the highest possible payoff under the constraint that player j best-replies to player i's strategy. The "flexible" player j cannot gain from a deviation, because the "blindness" of player i implies that player i's behavior remains the same regardless of player i's deviation, and, thus, player i cannot do better than best-replying to player i's strategy.

Proposition 7. *Game $G = (S, \pi)$ admits a BBE $\left(\left(\psi_i^*, Id \right), \left(s_i^*, s_j^* \right) \right)$ for each player i with the following properties: (1) ψ_i^* is blind, (2) s_i^* is an undominated Stackelberg strategy, and (3) $s_j^* = max_{s_j \in BR\left(s_i^*\right)} \left(\pi_i \left(s_i^*, s_j \right) \right)$. Moreover, $\left(\left(\psi_i^*, Id \right), \left(s_i^*, s_j^* \right) \right)$ is a strong BBE if $\{s_i^*\} = BR^{-1}\left(s_j^* \right)$.*

Proof. Let s_i^* be an undominated Stackelberg strategy of player i. Let

$$s_j^* = \text{argmax}_{s_j \in BR\left(s_i^*\right)} \left(\pi_i \left(s_i^*, s_j \right) \right).$$

Let $\hat{s}_j \in BR^{-1}(s_i^*)$ ($\{\hat{s}_j\} = BR^{-1}(s_i^*)$ with the additional assumption of the "moreover" part). We now show that $\left((\psi_i^* \equiv \hat{s}_j, Id), \left(s_i^*, s_j^*\right)\right)$ is a (strong) BBE. It is immediate that $\left(s_i^*, s_j^*\right) \in NE\left(G_{(\hat{s}_j, Id)}\right)$, and that both biased beliefs are monotone.

Next, observe that for any biased belief ψ_j' there is a plausible equilibrium (in any equilibrium) of the biased game $G_{\left(\hat{s}_j, \psi_j'\right)}$ in which player i plays s_i^*, and player j gains at most $\pi_j\left(s_i^*, s_j^*\right)$, which implies that the deviation to ψ_j' is not profitable to player j in this plausible equilibrium (in any equilibrium) of the new biased game.

If player i deviates to a biased belief ψ_i', then in any equilibrium of the biased game $G_{\left(\psi_i', Id\right)}$ player i plays some strategy s_i' and gains a payoff of at most $\max_{s_j' \in BR\left(s_i'\right)}\left(\pi_i\left(s_i', s_j'\right)\right)$, and this implies that player i's payoff is at most π_i^{Stac} , and that he cannot gain by deviating. This shows that $((\hat{s}_j, Id), (s_1^*, s_2^*))$ is a (strong) BBE. □

We demonstrate this class of equilibria in a Cournot competition.

Example 5 (*Well-behaved BBE that yields the Stackelberg outcome in a Cournot competition*)**.** Consider the symmetric Cournot game with linear demand in Example 1: $G = (S, \pi)$: $S_i = \mathbb{R}^+$ and $\pi_i(s_i, s_j) = s_i \cdot (1 - s_i - s_j)$ for each player i. Then $\left((0, I_d), \left(\frac{1}{2}, \frac{1}{4}\right)\right)$ is a strong well-behaved BBE that induces the Stackelberg outcome $\left(\frac{1}{2}, \frac{1}{4}\right)$, and yields the Stackelberg-leader payoff of $\frac{1}{8}$ to player 1 and yields the follower payoff of $\frac{1}{16}$ to player 2. This is because: (1) $\left(\frac{1}{2}, \frac{1}{4}\right) \in NE\left(G_{(0, I_d)}\right)$, (2) for any biased belief ψ_2', player 1 keeps playing $\frac{1}{2}$ and as a result player 2's payoff is at most $\frac{1}{16}$, and (3) for any biased belief ψ_1', player 2 will best-reply to player's 1 strategy, and thus player 1's payoff will be at most his Stackelberg payoff of $\frac{1}{8}$.

6.2 Folk Theorem Results

In this subsection we present various folk theorem results (i.e., general feasibility results) that show that relaxing either of the two requirements in the definition of a BBE (namely, monotonicity and ruling out implausible equilibria) yields little predictive power in various classes of games. Specifically, we show that in those games a strategy profile is a monotone weak

BBE outcome (resp., non-monotone strong BBE outcome) essentially iff it is (1) undominated, and (2) induces a payoff above the undominated minmax payoff.

6.2.1 Preliminary Definitions

We begin by defining the notions of monotone weak BBE, and of non-monotone strong BBE.

Definition 12. A weak BBE (ψ^*, s^*) is a *monotone weak BBE* if each biased belief ψ_i^* is monotone for each player i.
A *weak BBE* (ψ^*, s^*) is a *non-monotone strong BBE* if the inequality $\pi_i\left(s_i', s_j'\right) \leq \pi_i\left(s_i^*, s_j^*\right)$ holds for every player i, every biased belief ψ_i', and every strategy profile $\left(s_i', s_j'\right) \in NE\left(G_{\left(\psi_i', \psi_j^*\right)}\right)$.
Note that (1) a monotone weak BBE is a weakening of the notion of a BBE, which relaxes the requirement of ruling out implausible equilibria, and (2) a non-monotone strong BBE is a weakening of the notion of strong BBE, which relaxes the requirement of monotonicity.

6.2.2 Folk Theorem Result: Monotone Weak BBE in Finite Games

We say that a finite game G admits *best replies with full undominated support*, if, for each player i, there exists an undominated strategy $s_i \in S_i^U$ with a support that includes all undominated actions, i.e., $supp\left(s_i\right) = A_i \cap S_i^U$. Two classes of games that admit best replies with full undominated support are:

1. *All two-action games.* The reason for this is as follows. If player i has a dominant action, then, trivially, the dominant action a_i is an undominated strategy with a support that includes all undominated actions. If player i does not have a dominant action, then there must be a strategy of the opponent for which the player is indifferent between his two actions, which implies that there exists an undominated strategy with full support.

2. Any game with a totally mixed equilibrium (e.g., a rock-paper-scissors game).

Our next result focuses on finite games that admit best replies with full undominated support, and shows that in such games a strategy profile s^* is a monotone weak BBE outcome iff (I) s^* is undominated, and (II) the payoff of is s^* above the undominated minmax payoff.

The sketch of the proof is as follows. Each player has a blind belief that his opponent plays her part of the Nash equilibrium with full undominated support. This implies that each player is always indifferent between all undominated actions and, as such, can (1) play s_i^* on the equilibrium path, and (2) play a punishing strategy that guarantees the opponent a payoff of at most her undominated minmax payoff following any deviation of the opponent.

Proposition 8 (*Folk Theorem result for monotone weak BBE outcomes*)**.** *Let G be a finite game that admits best replies with full undominated support. Then the following two statements are equivalent:*

1. *Strategy profile (s_1^*, s_2^*) is a monotone weak BBE outcome.*
2. *Strategy profile (s_1^*, s_2^*) is (I) undominated and (II) $\pi_i(s_1^*, s_2^*) \geq M_i^U$.*

Proof. Proposition 3 implies that "1.⇒2." We now show that "2.⇒1." Assume that (s_1^*, s_2^*) is undominated, and $\pi_i(s_1^*, s_2^*) \geq M_i^U$. 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., $s_j^p = \text{argmin}_{s_j \in S_j^U}(\max_{s_i \in S_i} \pi_i(s_i, s_j))$. For each player j, let $s_j^e \in S_j^U$ be a best-reply strategy with full undominated support, i.e., $supp\left(s_j^e\right) = A_i \cap S_i^U$. For each player i, let $s_i^d \in BR^{-1}\left(s_j^e\right)$. The fact that $s_j^e \in BR\left(s_i^d\right)$ implies that $s_j^*, s_j^p \in \Delta\left(S_j^U\right) = \Delta\left(supp\left(s_j^e\right)\right) \subseteq BR\left(s_i^d\right)$.

We conclude by showing that $\left(\left(s_1^d, s_2^d\right), (s_1^*, s_2^*)\right)$ is a monotone weak BBE (in which both players have blind beliefs). It is immediate that $(s_1^*, s_2^*) \in NE\left(s_1^d, s_2^d\right)$. Next, observe that for any deviation of player i to a different biased belief ψ_i', there is a Nash equilibrium of the biased game $G_{\left(\psi_i', s_j^e\right)}$ in which player j plays s_j^p, and, as a result, player i obtains a payoff of at most M_i^U, which implies that the deviation is not profitable. Thus, (s_1^*, s_2^*) is a BBE outcome. □

Proposition 8 suggests that the notion of monotone weak BBE is too weak. The folk theorem result relies on the incumbents "discriminating" against deviators who have exactly the same perceived behavior as the rest of the population: the incumbents of population j "punish" deviators by playing s_j^p against them, while continuing to play s_j^* against the incumbents, even though both the deviators and the incumbents are perceived to behave the same (i.e., $\psi_j^*(s_i^e) = \psi_j^*(s_i^*)$).

Example 8 in Appendix A.3 demonstrates that the folk theorem result does not necessarily hold for games that do not admit best replies with full undominated support.

6.2.3 Folk Theorem Result: Non-Monotone Strong BBE in Interval Games

In this section we show a folk theorem result for strong BBE in a broad family of interval games in which each payoff function $\pi_i(s_i, s_j)$ is (1) strictly concave in s_i and (2) weakly convex in s_j. Examples of such games include Cournot competitions, price competitions with differentiated goods, public good games, and Tullock contests.

The following result shows that in this class of interval games, any undominated strategy profile (s_1^*, s_2^*) that induces each player a payoff strictly above the player's undominated minmax payoff can be implemented as an outcome of a strong BBE. Formally:

Proposition 9. *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 (s_1^*, s_2^*) is a non-monotone strong BBE outcome.*

The sketch of the proof is as follows (the formal proof is presented in Appendix F.9).

Each player j has a biased belief ψ_j^* that (I) distorts s_i^* into $BR^{-1}\left(s_j^*\right)$, and (II) distorts any s_i' that is not in a small neighborhood of s_i^*, to $BR^{-1}\left(s_j^p\right)$, where s_j^p is a "punishing" strategy that guarantees that player i obtains at most his undominated minmax payoff. Part (I) implies that (s_1^*, s_2^*) is an equilibrium of the biased game. Part (II) implies that following any deviation of player i to a different biased belief, if player i plays a strategy that is not in a small neighborhood of s_i^*, then player i loses from the deviation. Finally, the assumption that the payoff function $\pi_i(s_i, s_j)$ is convex in s_j implies that we can "complete" a continuous description of ψ_j^* for s_i' that are in a small neighborhood around s_i^*, such that a player cannot gain from deviating to playing strategies in this small neighborhood.

6.2.4 Discussion of the Folk Theorem Results

The results of this section show that the notion of weak BBE has little predictive power in the sense that, essentially, any undominated strategy profile with a payoff above the undominated minmax payoff is a weak BBE outcome. Moreover, we show that this multiplicity of BBE outcomes holds in large classes of games also when applying a refinement of monotonicity (Prop. 8), or when applying a refinement of strongness (Prop. 9). By contrast, in Section 5 we show that the combination of two plausible requirements, namely, monotonicity and ruling out implausible equilibria, allows us to achieve sharp predictions for the set of BBE outcomes in various interesting classes of games and for the set of biased beliefs that support these outcomes.

Our folk theorem results have similar properties to the famous folk theorem results for repeated games and sufficiently discounted players (see, e.g., Fudenberg and Maskin, 1986). This is so because it allows for implicit punishments similar to those used in repeated games in order to sustain equilibria. This is because our model assumes that when a player deviates to a different biased belief his opponent can react to the deviation and deter against it.

Observe that our result has somewhat stronger predictive power than the folk theorem result for repeated games, in the sense that the set of monotone weak BBE in one-shot finite games and the set of non-monotone strong BBE in one-shot interval games are each smaller than the set of subgame-perfect equilibria of repeated games between patient players. In particular, the following strategy profiles can be supported as the subgame-perfect equilibrium outcomes of a repeated game between patient players, but they cannot be the outcome of a weak BBE outcome of a one-shot game: (1) strategy profiles in which one of the players plays a strategy that is strictly undominated in the underlying (one-shot) game, and (2) strategy profiles in which some of the players obtain a payoff between the standard minmax payoff and the (higher) undominated minmax payoff.

In Appendix D we show that if one relaxes the assumption that the biased beliefs must be continuous, then one can obtain a folk theorem result in broader classes of games, namely, (1) in all finite games, and (2) in all interval games with strictly concave payoffs.

7 Conclusion

Decision makers' preferences and beliefs may intermingle. In strategic environments distorted beliefs can take the form of a self-serving commitment device. Our paper introduces a formal model for the persistence of such beliefs and proposes an equilibrium concept that supports them. Our analysis characterizes BBE in a variety of strategic environments, such as games with strategic complements and games with strategic substitutes. In particular, we show that agents present wishful thinking in all BBE in both of these common environments.

Our analysis here deals with simultaneous games of complete information, but the idea of strategically distorted beliefs may play an important role also in sequential games and in Bayesian games. In these frameworks, belief distortion may violate Bayesian updating, and our concept here can potentially offer a theoretical foundation for some of the cognitive biases relating to belief updating. It can also potentially identify the strategic environments in which these biases are likely to occur. We view this as an important research agenda that we intend to undertake in the future.

A different research track that might shed more light on strategic belief distortion is the experimental one. Laboratory experiments often conduct belief elicitation with the support of incentives for truthful revelation. Strong evidence for strategic belief bias in experimental games can be obtained by showing that players assign different beliefs to the behavior of their own counterpart in the game and to a person playing the same role with someone else. In general, our model predicts that beliefs about a third party's behavior are more aligned with reality than those involving one's counterpart in the game. Laboratory experiments can also test whether specific types of belief distortions (such as wishful thinking) arise in the strategic environments that are predicted by our model.

Finally, we point out that strategic beliefs may play an important role in the design of mechanisms and contracts. Belief distortions may destroy the desirable equilibrium outcomes that a standard mechanism aims to achieve. Mechanisms that either induce unbiased beliefs or adjust the rules of the game to account for possible belief biases are expected to perform better.

References

ACEMOGLU, D., AND M. YILDIZ (2001): “Evolution of perceptions and play,” mimeo.

ALGER, I., AND J. W. WEIBULL (2013): “Homo moralis, preference evolution under incomplete information and assortative matching,” *Econometrica*, 81(6), 2269–2302.

ATTANASI, G., AND R. NAGEL (2008): “A survey of psychological games: Theoretical findings and experimental evidence,” in *Games, Rationality and Behavior: Essays on Behavioral Game Theory and Experiments*, ed. by A. Innocenti, and P. Sbriglia, pp. 204–232. London: Palgrave Macmillan.

AUMANN, R., AND A. BRANDENBURGER (1995): “Epistemic conditions for Nash equilibrium,” *Econometrica*, pp. 1161–1180.

BABAD, E., AND Y. KATZ (1991): “Wishful thinking: Against all odds,” *Journal of Applied Social Psychology*, 21(23), 1921–1938.

BABCOCK, L., AND G. LOEWENSTEIN (1997): “Explaining bargaining impasse: The role of self-serving biases,” *Journal of Economic Perspectives*, 11(1), 109–126.

BAR-HILLEL, M., AND D. BUDESCU (1995): “The elusive wishful thinking effect,” *Thinking & Reasoning*, 1(1), 71–103.

BARBER, B. M., AND T. ODEAN (2001): “Boys will be boys: Gender, overconfidence, and common stock investment,” *The Quarterly Journal of Economics*, 116(1), 261–292.

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

——— (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144(1), 1–35.

Battigalli, P., M. Dufwenberg, and A. Smith (2015): "Frustration and anger in games," mimeo.

Battigalli, P., and D. Guaitoli (1997): "Conjectural equilibria and rationalizability in a game with incomplete information," in *Decisions, Games and Markets*, pp. 97–124. Berlin: Springer.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90(1), 166–193.

Budescu, D. V., and M. Bruderman (1995): "The relationship between the illusion of control and the desirability bias," *Journal of Behavioral Decision Making*, 8(2), 109–125.

Bulow, J. I., J. D. Geanakoplos, and P. D. Klemperer (1985): "Multimarket oligopoly: Strategic substitutes and complements," *Journal of Political Economy*, 93(3), 488–511.

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.

Dixit, A. (1987): "Strategic behavior in contests," *The American Economic Review*, 77(5), 891–898.

DOBSON, K., AND R.-L. FRANCHE (1989): "A conceptual and empirical review of the depressive realism hypothesis," *Canadian Journal of Behavioural Science*, 21(4), 419–433.

DUFWENBERG, M., AND W. GÜTH (1999): "Indirect Evolution vs. Strategic Delegation: A Comparison of Two Approaches to Explaining Economic Institutions," *European Journal of Political Economy*, 15(2), 281–295.

ESPONDA, I. (2013): "Rationalizable conjectural equilibrium: A framework for robust predictions," *Theoretical Economics*, 8(2), 467–501.

ESPONDA, I., AND D. POUZO (2016): "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84(3), 1093–1130.

EYSTER, E., AND M. RABIN (2005): "Cursed equilibrium," *Econometrica*, 73(5), 1623–1672.

FEHR, E., AND K. M. SCHMIDT (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114(3), 817–868.

FERSHTMAN, C., AND U. GNEEZY (2001): "Strategic delegation: An experiment," *RAND Journal of Economics*, 32(2), 352–368.

FERSHTMAN, C., K. L. JUDD, AND E. KALAI (1991): "Observable contracts: Strategic delegation and cooperation," *International Economic Review*, 32(3), 551–559.

FERSHTMAN, C., AND Y. WEISS (1998): "Social rewards, externalities and stable preferences," *Journal of Public Economics*, 70(1), 53–73.

FORBES, D. P. (2005): "Are some entrepreneurs more overconfident than others?," *Journal of Business Venturing*, 20(5), 623–640.

FRIEDMAN, D., AND N. SINGH (2009): "Equilibrium vengeance," *Games and Economic Behavior*, 66(2), 813–829.

Fudenberg, D., and D. K. Levine (1993): “Self-confirming equilibrium,” *Econometrica*, 61(3), 523–545.

Fudenberg, D., and E. Maskin (1986): “The folk theorem in repeated games with discounting or with incomplete information,” *Econometrica*, 54(3), 533–554.

Gannon, K., and H. Zhang (2017): “Evolutionary Justifications for Overconfidence,” mimeo.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): “Psychological games and sequential rationality,” *Games and Economic Behavior*, 1(1), 60–79.

Güth, W. (1995): “An evolutionary approach to explaining cooperative behavior by reciprocal incentives,” *International Journal of Game Theory*, 24(4), 323–344.

Güth, W., and S. Napel (2006): “Inequality aversion in a variety of games: An indirect evolutionary analysis,” *The Economic Journal*, 116, 1037–1056.

Güth, W., and M. Yaari (1992): “Explaining reciprocal behavior in simple strategic games: An evolutionary approach,” in *Explaining Process and Change: Approaches to Evolutionary Economics*, ed. by U. Witt, pp. 23–34. Ann Arbor: University of Michigan Press.

Guttman, J. M. (2003): “Repeated interaction and the evolution of preferences for reciprocity,” *The Economic Journal*, 113(489), 631–656.

Heifetz, A., E. Segev, et al. (2004): “The evolutionary role of toughness in bargaining,” *Games and Economic Behavior*, 49(1), 117–134.

Heifetz, A., C. Shannon, and Y. Spiegel (2007a): “The dynamic evolution of preferences,” *Economic Theory*, 32(2), 251–286.

——— (2007b): “What to maximize if you must,” *Journal of Economic Theory*, 133(1), 31–57.

Heller, Y. (2014): “Overconfidence and diversification,” *American Economic Journal: Microeconomics*, 6(1), 134–153.

Heller, Y., and E. Mohlin (2017): “Coevolution of deception and preferences: Darwin and Nash meet Machiavelli,” mimeo.

Heller, Y., and E. Winter (2016): “Rule rationality,” *International Economic Review*, 57(3), 997–1026.

Herold, F., and C. Kuzmics (2009): “Evolutionary stability of discrimination under observability,” *Games and Economic Behavior*, 67, 542–551.

Inoue, Y., Y. Tonooka, K. Yamada, and S. Kanba (2004): “Deficiency of theory of mind in patients with remitted mood disorder,” *Journal of Affective Disorders*, 82(3), 403–409.

Jehiel, P. (2005): “Analogy-based expectation equilibrium,” *Journal of Economic Theory*, 123(2), 81–104.

Koçkesen, L., E. A. Ok, and R. Sethi (2000): “Evolution of interdependent preferences in aggregative games,” *Games and Economic Behavior*, 31(2), 303–310.

Lord, C. G., L. Ross, and M. R. Lepper (1979): “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence,” *Journal of Personality and Social Psychology*, 37(11), 2098–2109.

Malmendier, U., and G. Tate (2005): “CEO overconfidence and corporate investment,” *Journal of Finance*, 60(6), 2661–2700.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, vol. 1. New York: Oxford university press.

Mayraz, G. (2013): “Wishful thinking,” Discussion paper, University of Melbourne.

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.

Nyarko, Y., and A. Schotter (2002): "An experimental study of belief learning using elicited beliefs," *Econometrica*, 70(3), 971–1005.

Palfrey, T. R., and S. W. Wang (2009): "On eliciting beliefs in strategic games," *Journal of Economic Behavior & Organization*, 71(2), 98–109.

Potters, J., and S. Suetens (2009): "Cooperation in experimental games of strategic complements and substitutes," *The Review of Economic Studies*, 76(3), 1125–1147.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *American Economic Review*, 83(5), 1281–1302.

Ross, L., and C. Anderson (1982): "Shortcomings in attribution processes: On the origins and maintenance of erroneous social judgments," in *Judgement under Uncertainty: Heuristics and Biases*, ed. by D. Kahiwmamm, P. Slovic, and A. Tversky, pp. 129–152. Cambridge: Cambridge University Press.

Rubinstein, A., and A. Wolinsky (1994): "Rationalizable conjectural equilibrium: Between Nash and rationalizability," *Games and Economic Behavior*, 6(2), 299–311.

Rustichini, A., and M. C. Villeval (2014): "Moral hypocrisy, power and social preferences," *Journal of Economic Behavior & Organization*, 107, 10–24.

Singh, N., and X. Vives (1984): "Price and quantity competition in a differentiated duopoly," *The RAND Journal of Economics*, 15(4), 546–554.

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.

Winter, E., I. Garcia-Jurado, and L. Mendez-Naya (2017): “Mental equilibrium and rational emotions,” *Management Science*, 63(5), 1302–1317.