THE EVOLUTION OF BIASED PERCEPTIONS*

Aviad Heifetz† and Yossi Spiegel‡

First version: December 1999
This version: October 2001

Abstract

We show that in a large class of pairwise strategic interactions, individuals who have perception biases and update their beliefs in a non Bayesian fashion will not only survive in the long run, but also prosper and take over the entire population. This result holds even when the interacting individuals do not always observe their rivals' types and even when individuals sometimes play against nature. This result suggests that in general there is no reason to believe that evolution will lead to a population of rational agents who revise their beliefs using Bayesian updating. To prove our results, we develop a simple methodology that allows us to study the long run evolution of types in the population starting from (almost) any initial distribution of types in a fully dynamic evolutionary context.

*Earlier drafts of this paper were circulated under the title: "On the Evolutionary Emergence of Optimism." We benefitted from comments of seminar participants at Berkeley, Caltech, Northwestern University, Stanford University, University of Chicago, UCLA, University of Illinois, the 2000 World congress of the Econometric Society in Seattle, the Southwest Economic Theory 2001 Conference in Caltech, the 2001 Spring Midwest Theory meetings in University of Wisconsin, and the 2001 Summer meetings of the Econometric Society in University of Maryland.

†The Division of Humanities and Social Sciences, California Instituite of Technology, mail code 228-77, Pasadena, CA 91125, and The Eitan Berglass School of Economics, Ramat Aviv, Tel Aviv University, 69978, Israel. email: heifetz@post.tau.ac.il

‡Recanati Graduate School of Business Administration, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel. email: spiegel@post.tau.ac.il.

1 Introduction

One of the cornerstones of economic analysis is the assumption that economic agents are rational. Among other things, the rationality assumption implies that agents form unbiased forecasts about their prospects and update these forecasts in a Bayesian fashion. This assumption however is at odds with a large body of evidence from the Psychology literature which shows that in general individuals tend to be overly optimistic or overconfident about their prospects even when faced with evidence to the contrary. For instance, most mentally healthy people were found to have somewhat unrealistically positive self-views, while the less mentally healthy perceive themselves more accurately (Taylor and Brown, 1988). In peer reviews, self-ratings of non-depressed individuals' were considerably more favorable than those given to them by others (Lewinsohn et al., 1980). Non-depressed individuals exhibited an illusion of control in a dice-throwing experiment (Fleming and Darley, 1986). Most individuals believe that their driving ability is above average (Svenson, 1981), and most young Americans know that half of U.S. marriages end in divorce, but they are confident that theirs will not (Lehman and Nisbett, 1985). Experimental work suggests that excess entry of new businesses that fail within several years may be due to overconfidence of entrepreneurs about their own ability in comparison with that of other entrepreneurs (Camerer and Lovallo, 1999). Interviews with new entrepreneurs revealed that their self-assessed chances of success were uncorrelated with objective predictors like education, prior experience, and initial capital, and were on average wildly off the mark (Cooper et al., 1988).[1] New life insurance agents who put an optimistic spin on setbacks by seeing them as flukes rather than viewing them as signs of incompetence sold more policies during their first year and were half as likely to quit (Seligman and Schulman, 1986). And, experimental work suggests that parties to legal disputes are reluctant to settle out of court because they hold overly optimistic beliefs about the merits of their case (Babcock and Loewenstein, 1997).

Informally, the rationality assumption is often justified by appealing to evolutionary arguments: individuals with biased perceptions who fail to maximize their true expected

[1] Over 80% of the interviewed enrepreneurs estimated their chances to succeed at 70% or better with 33% estimating them as certain. The same entrepreneurs estimated the mean chances of success for a business like their at merely 59%.

payoffs will not do as well as unbiased individuals and will therefore disappear from the population in the long run. In this paper we show that these arguments need not be right and in fact, there is a wide range of circumstances in which biased individuals with unrealistic beliefs not only survive, but also prosper and take over the entire population.

Our model considers a large population of individuals who are continuously and randomly matched in pairs to interact. The individuals differ from one another in the way they perceive their expected payoffs from these interactions: optimists overestimate the impact of their own actions, pessimists underestimate it, and only realists assess it correctly.[2] Since optimists or pessimists have biased perceptions about their own performance, they do not play a best response against their rivals' actions and therefore fail to maximize their own true (expected) payoffs. On the other hand, being recognized by rivals as having a biased perception may give individuals a strategic advantage by inducing rivals to alter their behavior in a favorable way. For moderate levels of perception biases, this strategic advantage outweighs the associated loss from failing to maximize the true payoff. Consequently, individuals with moderately biased perceptions do better on average than realists, and for a large class of interactions they take over the entire population in the long run. We show that this result holds even when the interacting individuals do not always observe their rivals' perceptions and even when individuals occasionally play against nature instead of being engaged in strategic interactions.

The idea that a biased objective function may confer a strategic advantage is well known at least since Schelling (1960) and was used extensively in many areas in economics including, Macroeconomics (e.g., Rogoff, 1985), International Trade (e.g., Brander and Spencer 1985, Eaton and Grossman 1986), Industrial Organization, (e.g., Brander and Lewis 1986, Fershtman and Judd 1987), and delegation (e.g., Green 1992, Fershtman, Judd and Kalai

[2]Troughout the paper when we will talk about optimism or pessimism we will only consider the personal belief about the impact of one's actions on one's payoff (but not the belief about the state of nature in general or about the impact of the actions of others on one's payoff). In a sense then, what we call optimism or pessimism can be interpreted as a self-serving bias or overconfidence (optimism) or deflated self-esteem or lack of self-confidence (pessimism). It should also be noted that in many ways the distinction between optimism and pessimism is largely semantic: optimism about the possibility of having a sunny day could also be viewed as pessimism about raining.

1991, Fershtman, and Kalai 1997, Katz 1991). In a similar vein, the literature on the evolution of preferences has shown that a population of "irrational" types who care about fairness (Güth and Yaari 1992, Huck and Oechssler 1998), are socially minded (Fershtman and Weiss 1997, 1998), altruistic (Bester and Güth 1998), spiteful (Possajennikov 2000, Bolle 2000), envious (Bergman and Bergman, 2000), concerned with relative success (Koçkesen, Ok, and Sethi 2000a, 2000b), or overconfident in financial markets (Kyle and Wang 1997, Benos 1998), may be evolutionary stable, i.e., immune to the appearance of few rational "mutants" with unbiased objectives.[3]

In this paper, we take these ideas one step further by characterizing the conditions under which biased perceptions would evolve in *a full-fledged, dynamic evolutionary context.* Instead of just showing that biased perceptions are evolutionary stable, we establish the conditions under which, starting from "almost" any initial distribution, the distribution of perception biases in the population will converge over time to a unit mass on some level of a biased perception.[4] To prove our result, we posit an artificial preliminary two-players game, in which the players simultaneously choose their types by committing to a degree of optimism or pessimism, knowing that subsequently they will play the Nash equilibrium of the game between the two committed types but will get the true expected payoffs associated with this equilibrium. We then prove that if this artificial "types" game is dominance solvable, then a regular payoff-monotonic dynamics will wipe out all serially dominated types, implying that the distribution of types (i.e., perception biases) in the population will converge in distribution to a unit mass at some unique type. The result that a regular payoff-monotonic dynamics wipes out all serially dominated strategies was already proved by Samuelson and

[3]Related ideas already appear in Frank (1987, 1988). The indirect evolutionary approach, where the preferences rather than strategies evolve over time, is employed also by Dekel and Scotchmer (1999), Dufwenberg and Güth (1999), Rogers (1994), Robson (1996a,b), Waldman (1994), and Vega-Redondo (1997). See also further references in the sequel.

[4]Other fully dynamic models of the evolution of preferences include Huck, Kirchsteiger, and Oechssler (1997) who deal with the emergence of an endowment effect in bargaining and Sandholm (2001) who deals with individual dispositions towards particular strategies. Apart from the general context, these papers differ from ours in that the dynamics in Huck, Kirchsteiger, and Oechssler (1997) is not shown to converge in the long-run, whereas Sandholm only studies 2x2 normal form games.

Zhang (1992) for the case where there are finitely many strategies. Since in our model there is a continuum of perception biases (i.e., strategies in the artificial types game), we first extend the Samuelson and Zhang result to the case where the set of strategies is a compact interval and the payoff function of each type is continuous, and then we use this extension to establish the convergence of the population to a unique perception bias. This methodology allows us to determine in a relatively straightforward way the long run population dynamics. Heifetz and Segev (2001) use the same methodology to study the dynamic emergence of tough characters in a salient class of bargaining mechanisms under asymmetric information. We believe that this methodology can be potentially useful in studying the evolution of additional types of individual biases.

Our analysis shows that the unique type of perception bias that emerges in the long run (optimism or pessimism) depends on two factors. First, it depends on whether the actions of one individual impose a positive or a negative externality on other individuals. Second, it depends on whether the actions of individuals are strategic substitutes or strategic complements. "Cautious" optimism emerges either when the individual actions impose negative externalities on others and the actions are strategic substitutes or when the individual actions impose positive externalities on others and the actions are strategic complements. In both cases, the aggressive behavior of optimists who overestimate the impact of their actions on their expected payoffs induces rivals to change their behavior in a favorable way. On the other hand, if the individual actions impose negative (positive) externalities on others and the actions are strategic complements (substitutes), the aggressive behavior of optimists induces rivals to change their behavior in an unfavorable way so over time the population converges to some (moderate) level of pessimism. Only when individuals do not engage in strategic interactions either because there are no externalities or because each individual has a dominant strategy (i.e., strategies are neither strategic substitutes nor strategic complements), does the population converges to realism.

Our results have both negative and positive implications. On the negative side, our results indicate that in the context of strategic interactions, one cannot appeal to evolutionary arguments to justify the rationality of players. This result stands in sharp contrast with Sandroni (2000) who shows that a market economy populated by agents who initially differ

in the accuracy of their predictions will nonetheless converge over time to a (competitive) rational expectations equilibrium as agents who make inaccurate predictions are driven out of the market. The reason for the difference between our result and Sandroni's is that in competitive markets, agents are anonymous and their individual impact on the equilibrium is negligible; hence, they cannot gain any strategic advantage by being biased. One can therefore conclude that while competitive markets favor agents who are able to make accurate predictions, strategic interactions favor individuals who have biased predictions.

On the positive side, our results provide an evolutionary explanation for well-documented perception biases like the *belief perseverance phenomenon,* which is the tendency to cling to one's beliefs in the face of contrary evidence, or the *confirmation bias,* which is the tendency to seek information that confirms one's own views and overlook evidence that disconfirm these views.[5] Individuals who hold biased perceptions about their prospects and fail to update their beliefs in a Bayesian fashion will gain a strategic advantage over rivals and hence, their frequency in the population will grow over time at the expense of realistic individual who use Bayesian updating. This suggests in turn that there is no reason to believe that over time, individuals will learn to update their beliefs in a Bayesian fashion.

The paper is organized as follows. Section 2 considers a simple example that exhibits most of the properties of the model. Section 3 discusses several interpretational issues and extensions. Section 4 explores general conditions on the payoff functions which are sufficient for our results to hold, and elaborates on the dynamic selection process and its properties. Section 5 concludes.

2 A simple example

In this section we consider a simple example that illustrates the main ideas in this paper. This example shows that under certain conditions, the evolution of perceptions will be such

[5] Reflecting on many experiments, Wason (1981) reports that once people have a wrong idea they "...evade facts, become inconsistent, or systematically defend themselves against the threat of new information relevant to the issue." For detailed discussion and review of some experimental evidence on the belief perseverance phenomenon and the confirmation bias, see for instance Ch. 10 in Myers (1998).

that over time, the distribution of perceptions in the population will converge in distribution to a unit mass at some (moderate) level of optimism. That is over time, individuals who are somewhat (unrealistically) optimistic about their prospects, will grow in number at the expense of other types of individuals and will eventually take over the entire population. This result can provide an evolutionary explanation for the large body of evidence mentioned in the Introduction on the prevalence of optimism and overconfidence.[6]

Consider a large population of individuals who are continuously and randomly matched in pairs to interact with one another. In every pairwise interaction, the matched individuals, $i = 1, 2$, choose actions $x^i \in \Re$. These actions can be thought of as the degree of effort or the level of investment the individuals put into the interaction.[7] Given a pair of actions x^1, x^2, the payoffs of the individuals are

$$\Pi^i(x^i, x^j) = (\alpha - bx^j - x^i)x^i, \qquad \text{for } i = 1, 2 \text{ and } j \neq i, \tag{2.1}$$

where $\alpha > 0$ and $-1 < b < 1$. Note that $\Pi^i_j = -bx^i$ and $\Pi^i_{ij} = b$ (subscripts are used to denote partial derivatives). Since in this example, x^i and x^j are both positive in the relevant range, it follows that when $b > 0$, the individuals impose negative externalities on one another (the larger is j's action, the lower is i's payoff), and moreover, actions are strategic substitutes in the sense of Bulow, Geanakoplos and Klemperer (1985) (the best-response functions are decreasing in the (x^1, x^2) space). In contrast, when $b < 0$, the individuals impose positive externalities on one another, and the actions are strategic complements.

Although the payoffs of all individuals are symmetric, individuals differ from one another in the way they perceive the interactions between them: Pessimistic types underestimate the value of α, optimistic types overestimate it, and only realistic types assess it correctly. Specifically, individual i conceives the value of α to be

$$\alpha^i = \alpha + \tau^i, \qquad \tau^i \in T = [\underline{\tau}, \overline{\tau}], \tag{2.2}$$

where $-\alpha \leq \underline{\tau} < 0 < \frac{\alpha}{5} < \overline{\tau}$. We assume that $-\alpha < \underline{\tau}$ in order to ensure that α^1 and α^2 are both positive and assume that $\frac{\alpha}{5} < \overline{\tau}$ in order to ensure that we get interior solutions.

[6] For an alternative exploration of optimism and self-confidence based on dynamic inconsistency, see Benabou and Tirole (1999a,b) and Brocas and Carrillo (1999).

[7] For some interpretations, it may be suitable to consider only non-negative actions. Our arguments continue to hold with such a restriction, though the analysis gets more involved.

The parameter τ^i is individual i's perception bias (we will also refer to this parameter as individual i's type). We will say that i is an *optimist* if $\tau^i > 0$, a *pessimist* if $\tau^i < 0$, and a *realist* if $\tau^i = 0$. An optimist overestimates the return to his actions for any given action taken by the other individual while a pessimist underestimates it. Substituting α^i for α in equation (2.1), the conceived payoffs of the individuals can be written as

$$U^i(x^i, x^j) = (\alpha - bx^j - x^i)x^i + \tau^i x^i, \qquad \text{for } i = 1, 2 \text{ and } j \neq i. \tag{2.3}$$

That is, the conceived payoff differs from the true payoff by an additive term $\tau^i x^i$ that depends on the individual's type, τ^i, and the individual's action, x^i. The unique Nash equilibrium of the game with these utility functions is $(\widehat{x}^1, \widehat{x}^2)$, where

$$\widehat{x}^1 = \frac{2(\alpha + \tau^1) - b(\alpha + \tau^2)}{4 - b^2}, \qquad \widehat{x}^2 = \frac{2(\alpha + \tau^2) - b(\alpha + \tau^1)}{4 - b^2}. \tag{2.4}$$

Substituting $\widehat{x}^1$ and $\widehat{x}^2$ into (2.1), the true equilibrium payoff of individual i is

$$f^i(\tau^i, \tau^j) \equiv \Pi^i(\widehat{x}^i, \widehat{x}^j) = \frac{\left(2(\alpha + \tau^i) - b(\alpha + \tau^j)\right)\left(2\alpha - (2 - b^2)\tau^i - b(\alpha + \tau^j)\right)}{\left(4 - b^2\right)^2}. \tag{2.5}$$

Now imagine that these true payoffs translate into fitness terms, so the instantaneous growth rates of the various types are monotonic in their average true payoffs from interacting with randomly matched individuals from the population. That is, the proportion of types with high current average payoffs tends to increase at the expense of types with low current average payoffs. The evolution of types therefore follows a regular, payoff monotonic dynamics, as shall be formally defined in Section 4. The mechanism by which the frequency of types evolves over time can be seen as either purely biological (types with higher payoffs have a higher ability to reproduce), as a process by which parents transmit their attitudes to life (i.e. their degree of "optimism") to their children via education or parental influence, or as a process by which more successful attitudes to life are imitated more often and increase in popularity.

Given this evolution of attitudes, what kind of perception will perform best and survive in the long run? To provide an answer, we consider a preliminary, artificial two-player "types game." In this game, each individual $i = 1, 2$ chooses a type $\tau^i \in T = [\underline{\tau}, \overline{\tau}]$ which in turn determines his assessment $\alpha^i = \alpha + \tau^i$ in the ensuing pairwise interaction. The payoff of individual i in the artificial types game is given by equation (2.5) above.

The best-response functions in the artificial types game is

$$BR^i(\tau^j) = \frac{b^2\left((2-b)\,\alpha - b\tau^j\right)}{4(2-b^2)}, \qquad \text{for } i = 1,2 \text{ and } j \neq i. \tag{2.6}$$

The best-response functions are downward sloping in the (τ^1, τ^2) space when $b > 0$, and upward sloping when $b < 0$. In other words, the types in this artificial types game are strategic substitutes (complements) whenever the actions are strategic substitutes (complements) in the original pairwise interaction. Moreover, since $|b| < 1$, the slope of the best-response functions BR^i is less than or equal to $\frac{1}{4}$ in absolute value. Therefore, the types game has a unique Nash equilibrium $(\widehat{\tau}, \widehat{\tau})$, where

$$\widehat{\tau} \equiv \frac{b^2}{4 + 2b - b^2}\alpha. \tag{2.7}$$

The assumption that $\frac{\alpha}{5} < \overline{\tau}$ ensures that $\widehat{\tau} < \overline{\tau}$.

Noting that the strategy set of each player is a one-dimensional compact interval, the payoff function of each player is continuous over the space of outcomes, twice differentiable, and strictly concave with respect to the player's strategy, and the slopes of the best-response functions are less than 1 in absolute value, it follows from Theorem 4 in Moulin (1984) that the types game can be also solved by an iterative process of elimination of dominated strategies. The unique outcome that survives this process is the Nash equilibrium, $(\widehat{\tau}, \widehat{\tau})$.[8] By Theorem 1 below, all other types which are serially dominated (i.e., do not survive iterative elimination process) are wiped out in a regular payoff-monotonic selection dynamics.. Hence,

Proposition 1: *For any initial distribution of types whose support is a compact interval that contains $\widehat{\tau}$, the distribution of types will converge in distribution under a payoff-monotonic selection dynamics to a unit mass at $\widehat{\tau}$; the density of all other types will converge to zero.*

[8] For instance, if $b = 1$ then player i's best-response function in the "types game" is $BR^i(\tau^j) = \frac{\alpha - \tau^j}{4}$. Since $\tau^1 \geq -\alpha$, it is therefore better for player 2 to be of type $\tau^2 = \frac{2\alpha}{4} = \frac{\alpha}{2}$ than have any higher type. Understanding this, having type $\tau^1 = \frac{\alpha - \frac{\alpha}{2}}{4} = \frac{\alpha}{8}$ is better for player 1 than having any lower type. But with this in mind, player 2 is better off having the type $\tau^2 = \frac{\alpha - \frac{\alpha}{8}}{4} = \frac{7\alpha}{32}$ than any higher type, and so on. The only type that survives this iterative elimination process is $\widehat{\tau} = \frac{\alpha}{5}$, which is the (unique) Nash equilibrium type.

Figure 1 depicts $\frac{\widehat{\tau}}{\alpha} = \frac{b^2}{4+2b-b^2}$ which is the optimism factor to which the population converges in the long-run (the percentage of exaggeration of α). As can be seen, the perceptions of individuals in the population converge to optimism (i.e., $\frac{\widehat{\tau}}{\alpha} > 0$) for any $b \neq 0$. At the extreme case where $b = 1$ (the case of large negative externalities and strong strategic substitutes), the degree of optimism is 20%. When $b = -1$ (the case of large positive externalities and strong strategic complements), the degree of optimism reaches 100%. Only in the knife-edge case where $b = 0$ (there are no strategic interactions between individuals since the payoff of one individual is independent of the actions of others), does the population converge to realism.

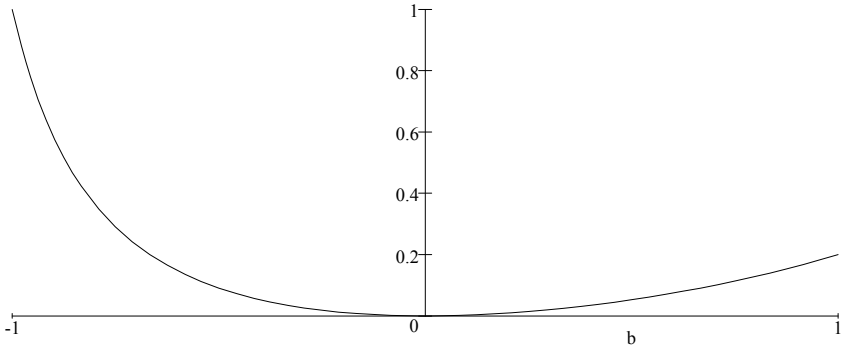

Figure 1: The optimism factor $\frac{\widehat{\tau}}{\alpha}$ at equilibrium as a function of b

The intuition underlying the evolution of optimistic perception in this example is as follows. Optimistic individuals play more aggressively than realists or pessimists and choose larger actions as they exaggerate the impact of their actions on their payoffs. When $b > 0$, actions are strategic substitutes, so the aggressive behavior of optimists induces rivals to play soft. Since $b > 0$ also implies that the actions of one individual impose a negative externality on others, the soft behavior of rivals benefits the optimistic individual. When $b < 0$, actions are strategic complements, so the aggressive behavior of optimists induces rivals to play aggressively as well. Since the actions of individuals impose positive externalities on others when $b < 0$, the aggressive behavior of rivals benefits the aggressive individual. Therefore, optimists gain a strategic advantage over rivals both when b is positive and negative. Of course, being aggressive is costly because an optimistic individual fails to play a best-response against his rival's action. Hence, wildly optimistic individuals do not do as well as more

moderately optimists, so on average, "cautiously" optimistic individuals fare better than individuals with other perceptions and therefore gradually take over the entire population.

3 Discussion

Before continuing with more general results, we sidestep to discuss several interpretational issues of the model. In addition, we show that our results continue to hold even if individuals are not always able to recognize each other's type and even if individuals occasionally play against nature rather than against other individuals.

3.1 Learning about α over time

The example in Section 2 immediately raises the following question: How come optimists do not realize that they have overestimated α once they observe their true payoffs? A possible answer can be due to two common and well-documented perception biases that were mentioned in the Introduction: the *belief perseverance phenomenon*, which is the tendency to cling to one's beliefs even after the basis on which they were formed has been discredited, and the *confirmation bias,* which is the tendency to seek information that confirms one's own views and overlook evidence that may disconfirm these views. These biases can explain why individuals can consistently overestimate the value of α and fail to learn about the true value of α over time. Our model can be used to explain why evolution does not eliminate the belief perseverance phenomenon and the confirmation bias: Individuals who hold optimistic beliefs about their prospects and fail to update them in a Bayesian fashion will gain a strategic advantage over other individuals and hence their frequency in the population will grow over time at the expense of realistic or pessimistic individuals. This suggests in turn that there is no reason to believe that individuals will learn over time to update their beliefs in a Bayesian fashion.[9]

[9] It is also worth noting that when two individuals are matched to interact, they "agree to disagree" about the value of α. This is because individuals in our model are fully aware of the confirmation bias, although each individual believes that he is not prone to this bias while others are. Therefore, each individual believes that his own estimate of α is unbiased, while his rival's estimates is biased and hence does not provide useful

In the rest of this subsection we make these arguments more precise. To this end, suppose that each period, α could either take on a high value, α_h, with probability p, or a low value, α_ℓ, with probability $1-p$, independently across periods. The true mean of α is $\overline{\alpha} \equiv p\alpha_h + (1-p)\alpha_\ell$. Individuals do not know the value of p, but can use past realizations of α (which can be inferred from their payoffs) to update their beliefs about p. Using $\overline{p}^i$ to denote individual i's posterior estimate of p, and noting from equation (2.3) that the payoff of individual i is linear in α^i, the expected perceived payoff functions are now given by equation (2.3) with $\alpha^i = \overline{p}^i\alpha_h + (1-\overline{p}^i)\alpha_\ell$.

Initially, individuals have in mind a prior distribution over values of p that contains the true value p in its support. Realists update their beliefs about p using the Bayes rule, so by the strong law of large numbers, their posterior distribution will converge over time in distribution to a point mass on p almost surely. As a result, the value of α^i if i is a realist will converge almost surely to the true mean $\overline{\alpha}$, so individual i will get closer and closer to maximizing his true expected payoff.

In contrast, optimists (pessimists) exhibit a confirmation bias and do not update their beliefs about p in a Bayesian fashion. Rather, they discard some constant proportion of low (high) realizations of α, attributing them the exceptional, non-systematic bad (good) circumstances, which render them irrelevant for updating. Consequently, if individual i is optimistic (pessimistic), his posterior estimate of p will converge almost surely to some point mass above (below) p. Hence α^i will converge almost surely to a value above (below) $\overline{\alpha}$. The higher the percentage of discarded realizations, the further α^i will get from $\overline{\alpha}$ and the larger will be the bias of individual i's perception. Since optimistic individuals who fail to update their beliefs in a Bayesian fashion gain a strategic advantage over realists and pessimists, their frequency in the population grows over time. Hence in the long run most individuals in the population will exhibit a confirmation bias.

information about the true value of α.

3.2 Types are imperfectly observed by opponents

A key assumption in Section 2 is that in every pairwise meeting, players recognize each other's type. One justification for this assumption could be that the players' types, τ^1 and τ^2, are conveyed through body language. Or it could be that each interaction consists of several rounds, in which the players' actions converge relatively quickly to the Nash equilibrium behavior even if at the outset players do not know each other's type (e.g., both players play their myopic best responses to each other's previous action or some average of previous actions). Clearly, if individual types are never observed before the pairwise interactions take place, players cannot gain a strategic advantage from being recognized as optimistic, so realists whose estimates of α are unbiased, would fare best. Consequently, as Dekel, Ely and Yilankaya (1998) and Ok and Vega-Redondo (2001) show (see also an example in Possajenikov 1999), the evolutionary stable distribution of the population is such that all individuals play their (true) Nash equilibrium strategy.

In this subsection we show that the qualitative results of Section 2 continue to hold even if players cannot always recognize each other's type. So long as there is a positive probability that types will be mutually observed, the population will still converge over time to a unique level of ("cautious") optimism, although the resulting perception bias is smaller than in the case where types are perfectly observed. To establish this result, suppose that unlike in Section 2, now players can observe each other's type only with probability $(1-p)$. With probability p, both players cannot observe each other's types.

When types are mutually observable, the equilibrium is as in Proposition 1. When types are not observable, we need to look for a Bayesian Nash equilibrium in which each player forms a belief about his opponent's action and plays a best-response given this belief. To characterize this equilibrium, let $\overline{x}$ be the average action in the population. Then, the perceived average payoff of player i whose type is τ^i when taking action x^i and facing a player with an unknown type is given by:

$$U^i(x^i, \overline{x}) = (\alpha + \tau^i - b\overline{x} - x^i)x^i. \tag{3.1}$$

The best-response of player i against $\overline{x}$ is:

$$BR^i(\overline{x}) = \frac{\alpha + \tau^i - b\overline{x}}{2}. \tag{3.2}$$

On the equilibrium path, player i's belief about $\overline{x}$ must be correct. Taking expectations on both sides of equation (3.2), using θ to denote the average type in the (current) population, and solving for $\overline{x}$ yields:

$$\overline{x} = \frac{\alpha + \theta}{2 + b}. \tag{3.3}$$

That is, when player i cannot observe his opponent's type, he (correctly) anticipates that given θ, his opponent will play on average $\overline{x}$. Substituting for $\overline{x}$ in (3.2) reveals that when player i meets player j and both cannot observe each other's types, their respective equilibrium actions are $\alpha + \tau^i - b\frac{\alpha+\theta}{2+b}$ and $\alpha + \tau^j - b\frac{\alpha+\theta}{2+b}$. Substituting these actions in equation (2.1), the true payoff of player i in equilibrium is given by

$$\left(\alpha - b\frac{\alpha + \tau^j - b\frac{\alpha+\theta}{2+b}}{2} - \frac{\alpha + \tau^i - b\frac{\alpha+\theta}{2+b}}{2}\right)\left(\frac{\alpha + \tau^i - b\frac{\alpha+\theta}{2+b}}{2}\right). \tag{3.4}$$

To determine the evolution of types in the population, we use the same methodology as in Section 2. That is, we consider a preliminary, artificial two-player "types" game in which players i and j choose their respective types, τ^i and τ^j, where $\tau^i, \tau^j \geq -\alpha$, and the expected payoff of type τ^i, given the probability that types will not be mutually observed, p, and given the current average type in the population, θ, is given by

$$\begin{aligned} f^i(\tau^i, \tau^j; p, \theta) &= (1-p)\frac{\left(2(\alpha+\tau^i) - b(\alpha+\tau^j)\right)\left(2\alpha - (2-b^2)\tau^i - b(\alpha+\tau^j)\right)}{\left(4-b^2\right)^2} \\ &\quad + p\left(\alpha - b\frac{\alpha + \tau^j - b\frac{\alpha+\theta}{2+b}}{2} - \frac{\alpha + \tau^i - b\frac{\alpha+\theta}{2+b}}{2}\right)\left(\frac{\alpha + \tau^i - b\frac{\alpha+\theta}{2+b}}{2}\right). \end{aligned} \tag{3.5}$$

With probability $1-p$ types are observed and individual i's payoff is as in Section 2, whereas with probability p types are not observed and i's payoff is given by the expression in (3.4). The best-response function of type τ^i against type τ^j in the artificial "types" game is:

$$\begin{aligned} BR^i(\tau^j; p, \theta) &= \frac{4\alpha b^2\left(2-b\right)\left(1-p\right)}{2\left(16 - 8b^2 + pb^4\right)} + \frac{bp\left(2-b\right)\left(4b + 8 - b^3 - 2b^2\right)}{2\left(16 - 8b^2 + pb^4\right)}\theta \\ &\quad - \frac{b\left(pb^4 + 4b^2 - 12pb^2 + 16p\right)}{2\left(16 - 8b^2 + pb^4\right)}\tau^j. \end{aligned} \tag{3.6}$$

In what follows we prove that the population evolves over time to a stable monomorphic type. To this end, let G_t be the distribution of types in the population at time $t \geq 0$ on the support $T = [\underline{\tau}, \overline{\tau}]$ and let

$$\begin{aligned} \overleftarrow{\tau} &= \inf\left\{\tau > \tau^*(p) : \exists V_\tau \ni \tau, \quad V_\tau \text{ open, s.t. } \lim_{t\to\infty} G_t(V_\tau) = 0\right\}, \\ \underrightarrow{\tau} &= \sup\left\{\tau < \tau^*(p) : \exists V_\tau \ni \tau, \quad V_\tau \text{ open, s.t. } \lim_{t\to\infty} G_t(V_\tau) = 0\right\}. \end{aligned} \tag{3.7}$$

Then we can prove the following result:

Proposition 2: *Suppose that players can observe each other's type only with probability* $(1-p)$. *Then under a payoff-monotonic selection dynamics, the distribution of types in the population converges in distribution to a unit mass at* $\tau^*(p)$, *such that*

$$\overleftarrow{\tau} = \underrightarrow{\tau} = \tau^*(p) = BR^i(\tau^*(p); p, \tau^*(p)) = \frac{2b^2(1-p)}{8+4b-2b^2-pb^3}\alpha.$$

Proof of Proposition 2: *See the Appendix.*

The idea behind the proof is as follows. The artificial types game played at each point in time depends on the current average type θ. Hence, we cannot prove the convergence result as in Section 2 and need to use a more involved methodology. Yet, once it is determined that, irrespective of the value of θ, types outside an interval $[\tau_\ell, \tau_h]$ are serially dominated and hence asymptotically extinct under a payoff-monotonic selection dynamics, the average type θ will eventually settle in the interval $[\tau_\ell - \varepsilon, \tau_h + \varepsilon]$, where ε is positive and small.[10] The fact that $\theta \in [\tau_\ell - \varepsilon, \tau_h + \varepsilon]$ enables us to show that even more types are serially dominated and thus that types outside some smaller interval $[\tau'_\ell, \tau'_h] \subset [\tau_\ell, \tau_h]$ are also asymptotically extinct. The crux of the argument is in showing that it is impossible for this iterative process to halt with an interval $[\tau'_\ell, \tau'_h]$ of positive length.

Since $\tau^*(0) = \widehat{\tau}$, Proposition 2 implies that when $p = 0$ (full observability), the population converges to the same type as in Proposition 1. On the other hand, $\tau^*(1) = 0$, implying that when $p = 1$ (non-observability), the population converges to realism. This

[10] In general, θ will settle in the interior of the interval $[\tau_\ell, \tau_h]$. However, a-priori we cannot rule out the possibility that θ will either approach τ_ℓ from below or τ_h from above and will therefore stay outside the interval $[\tau_\ell, \tau_h]$. Therefore, we show that θ will settle in the (larger) interavl $[\tau_\ell - \varepsilon, \tau_h + \varepsilon]$.

is because optimists cannot gain a strategic advantage if types are never observed before actions are taken. For levels of p between 0 and 1, $\tau^*(p)$ decreases continuously from $\widehat{\tau}$ to 0. That is, the higher is p, the lower is the level of optimism to which the population will converge over time.

3.3 Individuals occasionally play against nature

The main point of the analysis so far was that in the context of strategic interactions, individuals with biased perceptions gain a strategic advantage over rivals and as a result, the population converges over time to a biased monomorphic type. In this subsection we show that this argument continues to hold even if occasionally, individuals play against nature instead of being engaged in strategic interactions against other individuals. To capture this idea in a simple way, we shall assume that with probability $1-\rho$, the situation is exactly as in Section 2. With probability ρ, however, the situation is such that $b=0$. In the latter case, there is no strategic interaction between individuals (the payoff of i is independent of j's action), so individuals cannot gain a strategic advantage by being optimistic. The expected payoff of type τ^i when facing type τ^j and when $b=0$ with probability ρ is given by:

$$\begin{aligned} f^i(\tau^i,\tau^j;\rho) &= (1-\rho)\frac{\left(2(\alpha+\tau^i)-b(\alpha+\tau^j)\right)\left(2\alpha-(2-b^2)\tau^i-b(\alpha+\tau^j)\right)}{\left(4-b^2\right)^2} \\ &\quad +\rho\frac{\left((2-b)\,\alpha+2\tau^i\right)\left((2-b)\,\alpha-2\tau^i\right)}{4^2}. \end{aligned} \tag{3.8}$$

The expression multiplied by $1-\rho$ is the payoff in (2.5). The expression multiplied by ρ is the payoff in (2.5) evaluated at $b=0$.

The best-response functions in the artificial two-players "types game" are given by

$$BR^i(\tau^j;\rho)=\frac{2(1-\rho)b^2\left(2\alpha-b(\alpha+\tau^j)\right)}{16-8b^2+\rho b^4}, \qquad \text{for } i=1,2 \text{ and } j\neq i. \tag{3.9}$$

As before, the best-response functions are downward sloping in the (τ^1,τ^2) space when $b>0$, and upward sloping when $b<0$. Moreover, since $|b|<1$, the slope of the best-response functions is less than or equal to $\frac{1}{4}$ in absolute value. Therefore, the artificial "types game" can be solved by a process of iterative elimination of dominated strategies. The unique type

that survives this process is given by the intersection of $BR^1(\tau^2;\rho)$ and $BR^2(\tau^1;\rho)$ in the (τ^1,τ^2) space which is

$$\widehat{\tau}(\rho) \equiv \frac{2(1-\rho)b^2}{2\left(4+2b-b^2\right)-\rho b^3}\alpha. \tag{3.10}$$

The type $\widehat{\tau}(\rho)$ is strictly decreasing with ρ. Hence, the more often individuals have to play against nature (i.e., the closer is ρ to 1), the smaller is the perception bias to which the population will converge over time. At the extreme when $\rho = 1$ (individuals play only against nature but not against one another), $\widehat{\tau}(1) = 0$ so the population converges to realism. At the other extreme where $\rho = 0$ (individuals never play against nature), $\widehat{\tau}(0) = \widehat{\tau}$, where $\widehat{\tau}$ is given by equation (2.7), so the situation is as in Section 2. The main point however is that so long as there are strategic interaction between individuals (i.e., $\rho < 1$) the qualitative results from Section 2 continue to hold except that now, the degree of bias is smaller since $\widehat{\tau}(\rho) < \widehat{\tau}$ for all $\rho > 0$.

Proposition 3: *Suppose that players engage in strategic interaction only with probability* $(1-\rho)$. *Then under a payoff-monotonic selection dynamics, the distribution of types in the population converges in distribution to a unit mass at* $\widehat{\tau}(\rho)$, *where* $\widehat{\tau}(\rho)$ *decreases from* $\widehat{\tau}$ *when* $\rho = 0$ *to* 0 *when* $\rho = 1$.

4 A general analysis

In the example considered in Section 2, the conceived payoff of individual i from interacting with individual j was of the form

$$U^i(x^i,x^j) = \Pi(x^i,x^j) + \tau^i x^i, \qquad i = 1,2, \text{and} \;\; j \neq i, \tag{4.1}$$

where $\Pi(x^i,x^j) = (\alpha - bx^j - x^i)x^i$, and τ^i represented individual i's perception bias or type. Optimistic types had a positive τ^i, and thus overestimated their personal benefit from their own action, x^i, while pessimistic types had a negative τ^i and thus underestimated their personal benefit from their own actions.

In this section we continue to consider additive linear perception biases as in (4.1) but allow for a much broader class of (true) payoff functions $\Pi^i \equiv \Pi(x^i,x^j)$. Throughout we shall

assume that Π^i is thrice continuously differentiable and will denote its partial derivatives by

$$\Pi^i_i \equiv \frac{\partial \Pi(x^i, x^j)}{\partial x^i}, \quad \Pi^i_j \equiv \frac{\partial \Pi(x^i, x^j)}{\partial x^j}, \quad \Pi^i_{ii} \equiv \frac{\partial^2 \Pi(x^i, x^j)}{(\partial x^i)^2}, \quad \textit{and} \quad \Pi^i_{ij} \equiv \frac{\partial^2 \Pi(x^i, x^j)}{\partial x^i \partial x^j}.$$

In addition, we shall assume that Π^i has the following properties:

Property 1: Π^i is strictly concave in x^i, i.e., $\Pi^i_{ii} < 0$.

Property 2: Π^i is such that $\left|\frac{\Pi^i_{ij}}{\Pi^i_{ii}}\right| < 1 - \varepsilon$, for some $\varepsilon > 0$.

Property 1 ensures that i's best-response against j's action, $BR^i(x^j)$, is implicitly defined by the first order condition

$$U^i_i(x^i, x^j) = \Pi^i_i + \tau^i = 0. \tag{4.2}$$

Property 2 ensures that the slope $\frac{\mathrm{d}BR^i(x^j)}{\mathrm{d}x^j}$ of i's best-response function is uniformly smaller than 1 in absolute value. This implies in turn that every pair of best-response functions intersect exactly once in the actions space, so every pairwise interaction has a unique Nash equilibrium.[11]

Let $(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j))$ be the (unique) Nash equilibrium in the interaction between individuals i and j given their types. The equilibrium strategies are implicitly defined by the following two equations:

$$\begin{aligned} \Pi_i(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) + \tau^i &= 0, \\ \Pi_j(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) + \tau^j &= 0. \end{aligned} \tag{4.3}$$

To study the evolution of perception biases, we once again consider an artificial two-player "types game" in which each player i selects a type $\tau^i \in T = [\underline{\tau}, \tau]$, where $\underline{\tau} < 0 < \overline{\tau}$. The payoff of player i in the "types game" is given by

$$f^i(\tau^i, \tau^j) \equiv \Pi(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)), \tag{4.4}$$

[11] The uniformity requirement is needed to guarantee that the best-respone functions are not mutually asymptotic without intersecting. Actually, property 2 ensures that a myopic best-reply process in a repetition of the game (in which each individual plays a best reply to the previous action of the opponent) converges to the unique Nash equilibrium. This may justify the assumption that (when each interaction consists of several rounds) individuals essentially play the unique Nash equilibrium even if initially they do not recognize each other's type.

which is the true equilibrium payoff in the interaction between i and j. The uniqueness of the Nash equilibrium in every pairwise interaction implies that $f^i(\tau^i, \tau^j)$ is well-defined. Since all individuals have the same true payoff function $\Pi(\cdot, \cdot)$ (individuals differ only with respect to the value of τ), the payoffs in the types game are symmetric in the sense that $f^i(\tau^i, \tau^j) = f^j(\tau^j, \tau^i)$. We shall now assume that the payoff function $f^i \equiv f^i(\tau^i, \tau^j)$ has the following properties:

Property 3: f^i is twice differentiable and strictly concave in τ^i, i.e., $f^i_{ii} < 0$.

Property 4: f^i is such that $|f^i_{ii}| < \left|f^i_{ij}\right|$.

Property 3 ensures that the types game is well-behaved. Property 4 ensures that the types game has a unique Nash equilibrium because it implies that the slope of each player's best-response function in the types game is less than 1 in absolute value.[12]

Remark: The set of action games that satisfy Properties 1-4 constitute an open set in the space $C^3(R^2_+)$ (the space of thrice continuously differentiable functions $\Pi : R^2_+ \to R$, with the minimal topology in which Π_n converges to Π if and only if $(\Pi_n - \Pi)$ and each of its first, second, and third derivatives converge to zero uniformly on compact sets on R^2_+), because they are defined using finitely many strict inequalities which involve continuous functions of up to third-order derivatives of Π^i.[13] Thus, the family of quadratic payoff functions studied in Section 2, which clearly satisfies Properties 1-4, is not exceptional. Other payoff functions whose derivatives up to the third-order are not "too far" from one of those quadratic functions will lead to the same kind of evolution of perception biases.

[12] We state Properties 3 and 4 in terms of the payoff function in the types game, f^i, because stated in terms of the payoff function in the actions game, Π^i, these properties involve complex expressions that include third-order derivatives of Π^i, which do not have an intuitive economic interpretation.

[13] More precisely, for every fixed $\varepsilon > 0$ in Property 2 there are finitely many strict inequalities involved in the definition of this set of payoff functions – denote it V_ε – which is therefore open. The actual set of payoff functions Π which obey all the properties is the union of these open V_ε over all the positive ε, which is open as the union of open sets.

4.1 Optimism and Pessimism

To study the properties of the "types game" further, let $\widehat{x}_i^i$ and $\widehat{x}_j^i$ be the partial derivatives of individual i's equilibrium action in the interaction with individual j with respect to his own type, τ^i, and j's type, τ^j. By differentiating (4.3) with respect to τ^i and τ^j and using Cramer's rule, we obtain the following comparative statics:

$$\widehat{x}_i^i = -\frac{\Pi_{jj}^j}{\det J}, \quad \widehat{x}_j^i = \frac{\Pi_{ji}^i}{\det J}, \tag{4.5}$$

where

$$J = \begin{pmatrix} \Pi_{ii}^i & \Pi_{ij}^i \\ \Pi_{ji}^j & \Pi_{jj}^j \end{pmatrix}.$$

Property 2 implies $\det J > 0$. Hence (4.5) and property 1 imply that $\widehat{x}_i^i > 0$ and $\widehat{x}_i^j \gtreqless 0$ as $\Pi_{ij}^j \gtreqless 0$. That is, as τ^i increases and individual i becomes more optimistic, he becomes more aggressive in the sense that his equilibrium action increases, while his opponent becomes more aggressive if actions are strategic complements ($\Pi_{ij}^j > 0$) but softer if actions are strategic substitutes ($\Pi_{ij}^j < 0$).

Given Property 3 and using (4.5), an interior Nash equilibrium in the "types game" is defined implicitly by the following pair of equations:

$$\begin{aligned} f_i^i(\tau^i, \tau^j) &= \frac{-\Pi_i^i \Pi_{jj}^j + \Pi_j^i \Pi_{ji}^j}{\det J} = 0, \\ f_j^j(\tau^i, \tau^j) &= \frac{-\Pi_j^j \Pi_{ii}^i + \Pi_i^j \Pi_{ij}^i}{\det J} = 0, \end{aligned} \tag{4.6}$$

where the partial derivatives of the actions game are evaluated at the Nash equilibrium actions $(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j))$. Since the payoff functions in the types game are symmetric, the Nash equilibrium in the types game is also symmetric and given by $(\widehat{\tau}, \widehat{\tau})$, where $\widehat{\tau}$ is defined implicitly by the equation $f_i^i(\widehat{\tau}, \widehat{\tau}) \equiv 0$.

To interpret the equilibrium conditions in the types game, note that (4.6) implies that at an interior Nash equilibrium of the type game we have

$$-\frac{\Pi_i^i}{\Pi_j^i} = -\frac{\Pi_{ji}^j}{\Pi_{jj}^j}. \tag{4.7}$$

The left side of (4.7) represents the slope of an iso-payoff curve of individual i in the underlying actions game in the (x^i, x^j) space, while the right side of (4.7) represents the slope of the best response function of individual j in the (x^i, x^j) space. Thus, equation (4.7) says that individual i chooses τ^i optimally in the types game by selecting the "highest" true iso-payoff curve, taking as given the best response function of the rival in the action game. The first order condition for this constrained maximization problem requires that, holding τ^j constant, the iso-payoff curve of individual i will be tangent to the best response function of individual j.[14]

Next, we study the conditions under which the symmetric Nash equilibrium of the types game is such that $\widehat{\tau} > 0$ (i.e., individuals "choose" to become optimistic).

Lemma 1: *The Nash equilibrium in the artificial "types game" is such that* $\widehat{\tau} > 0$ *if* Π^i_j *and* Π^i_{ij} *have the same sign,* $\widehat{\tau} < 0$ *if* Π^i_j *and* Π^i_{ij} *have the opposite sign, and* $\widehat{\tau} = 0$ *if* $\Pi^i_j \Pi^j_{ji} = 0$.

Proof: Substituting from (4.7) into (4.3) and rearranging terms yields

$$\widehat{\tau} = -\frac{\Pi^i_j \Pi^j_{ji}}{\Pi^j_{jj}}, \tag{4.8}$$

where the right hand side is evaluated at the symmetric Nash equilibrium $(\widehat{x}(\widehat{\tau}, \widehat{\tau}), \widehat{x}(\widehat{\tau}, \widehat{\tau}))$ in the interaction between two individuals who both have the type $\widehat{\tau}$. Since $\Pi^j_{jj} < 0$ by Property 1 and since $\Pi^j_{ji} = \Pi^i_{ij}$ by the symmetry of the actions game, it follows that $\widehat{\tau} > 0$ if and only if $\Pi^i_j \Pi^i_{ij} > 0$. ■

To interpret Lemma 1, note that Π^i_j captures the externality that individual j's action imposes on individual i's payoff in a pairwise interaction between them (the externality is positive if $\Pi^i_j > 0$, and negative if $\Pi^i_j < 0$), and the sign of Π^i_{ij} determines whether the actions are strategic substitutes ($\Pi^i_{ij} < 0$) or strategic of complements ($\Pi^i_{ij} > 0$). Hence Lemma 1 shows that in a Nash equilibrium of the types game, players "choose" to become optimistic (pessimistic) either if they impose negative (positive) externalities on one another and their actions are strategic substitutes (complements) or if they impose positive (negative) external-

[14] This is exactly like the behavior of a leader in a Stackelberg duopoly model who chooses the level of output at which its iso-profit curve is tangent to the best response function of the follower.

ities on one another and their actions are strategic complements (substitutes).[15] Intuitively, an optimist who overestimates the return to his actions, behaves more "aggressively" than a realist and chooses a higher level of action. When actions are strategic substitutes and the actions of one individual impose a negative externality on the payoffs of rivals, this aggressive behavior triggers a favorable soft behavior from the rival. Examples for such interactions include the tragedy of the commons (the joint use of congested common resources) and some tournaments.[16] When actions are strategic complements and the actions of one individual impose a positive externality on the payoff of rivals, the aggressive behavior of the optimist triggers a favorable aggressive behavior from rivals. Examples for this kind of interaction include the Bertrand duopoly model with differentiated products and the Cournot duopoly model with complementary products. In either case, the aggressive behavior associated with being an optimist gives individuals a strategic advantage; not surprisingly then, in a Nash equilibrium of the "types game", both players will choose to become optimistic. The reason why the players will only choose a moderate level of optimism is that the strategic advantage from being "wildly" optimistic is outweighed by the associated loss from having a biased perception and taking suboptimal actions. The intuition why players "choose" to become pessimistic when Π^i_j and Π^i_{ij} have the opposite signs is exactly the opposite. Games in which Π^i_j and Π^i_{ij} have the same sign include arms races and voluntary contribution to public goods games.[17]

The following lemma will be useful for the sequel:

[15] Fershtman and Weiss (1998) show that when Π^i_j and Π^i_{ij} have the same sign, social mindedness (enjoying doing what is socialy highly considered) is evolutionary stable.

[16] For instance, in Lazear and Rosen (1981), two individuals compete for a prize, w. Each individual expands effort, μ_i $(i = 1, 2)$, to produce an output $q_i = \mu_i + \epsilon_i$, where ϵ_i is a random shock. The individual with the higher output wins the prize. If the random shocks are independently drawn from the same exponential distribution, $F(\epsilon_i) = 1 - e^{-\lambda\epsilon_i}$, the expected payoff of individual i is $\pi_i = \frac{we^{-\lambda(\mu_j - \mu_i)}}{2} - C(\mu_i)$, where $C(\mu_i)$ is the disutility of effort which is increasing and convex. It can be verified that in this model, the efforts are strategic substitutes and the effort of one individual lowers the expected payoff of the other individual so Π^i_j and Π^i_{ij} are both negative.

[17] In a typical arms race game, $\Pi^i = V(x^i - x^j) - C(x^i)$,with $V'(.) > 0 > V"(.)$, $C'(.) > 0$, and $C"(.) \geq 0$, so $\Pi^i_j < 0 < \Pi^i_{ij}$, while in typical voluntary contribution to public goods games, $\Pi^i = V(x^i + x^j) - C(x^i)$, with $V'(.) > 0 > V"(.)$, and $C'(.) > 0$ and $C"(.) \geq 0$, so that $\Pi^i_j > 0 > \Pi^i_{ij}$.

Lemma 2: *The artificial "types game" is strictly dominance solvable. The unique outcome that survives the process of iterated elimination of strictly dominated strategies is the symmetric Nash equilibrium* $(\widehat{\tau}, \widehat{\tau})$.

Proof: To prove the lemma we invoke Theorem 4 in Moulin (1984) that provides sufficient conditions for normal form games to be dominance solvable. In the present context, these conditions are:

(i) The strategy set of each player is a one-dimensional compact interval.

(ii) The payoff function of each player is continuous over the space of outcomes, twice differentiable, and strictly concave with respect to the player's strategy.

(iii) The slope of each players' best-response functions is less than 1 in absolute value.

Condition (i) is satisfied in the types game because the set of possible types for each player is a (compact) interval $T = [\underline{\tau}, \overline{\tau}]$. Properties 3 and 4 ensure that conditions (ii) and (iii) are satisfied. Hence, the types game is dominance solvable. ■

4.2 The Evolutionary dynamics of perception biases

We now turn to the way the population of types evolves over time. To this end, let G_t be the distribution of types in the population at time $t \geq 0$ on the support $T = [\underline{\tau}, \overline{\tau}]$. We assume that G_t evolves according to a payoff-monotonic selection dynamics, where types with higher average payoffs have higher growth rates. Specifically, we shall define the growth rate of types as follows:

Definition. A continuous growth-rate function $g : T \times \Delta(T) \to R$ is payoff-monotonic and regular if for every $G \in \Delta(T)$, the following holds:

(i) Higher payoffs are associated with higher growth rates:

$$g(\tau^i, G) > g(\widetilde{\tau}^i, G) \quad \Longleftrightarrow \quad f(\tau^i, G) > f(\widetilde{\tau}^i, G). \tag{4.9}$$

(ii) g is a mass preserving spread of G (i.e., the densities of the various types of individuals in the population sum up to 1):

$$\int_T g(\cdot, G)\mathrm{d}G(\cdot) = 0. \tag{4.10}$$

(iii) g can be extended to the domain $T \times X$, where X is the set of signed measures G with variational norm smaller than 2, and on this extended domain, g is bounded and Lipschitz continuous:

$$\begin{aligned} \sup_{\tau^i \in T, G_t \in X} \left| g(\tau^i, G) \right| &< \infty, \\ \sup_{\tau^i \in T} \left| g(\tau^i, G) - g(\tau^i, \widetilde{G}) \right| &< K \left\| G_t - \widetilde{G} \right\|, \qquad G, \widetilde{G} \in X, \end{aligned} \tag{4.11}$$

for some constant K, where $\|G\| = \sup_{|h| \leq 1} \left| \int_T h\mathrm{d}G \right|$ is the variational norm on signed measures.

Oechssler and Riedel (2001, Lemma 3) prove that property (iii) guarantees that the mapping $G \to \int_T g(\cdot, G)\mathrm{d}G$ is bounded and Lipschitz continuous in the variational norm, which implies that the differential equation in the space of distributions $\Delta(T)$ defined by

$$\dot{G_t}(S) = \int_S g(\cdot, G_t)\mathrm{d}G_t(\cdot), \qquad S \subseteq T, \tag{4.12}$$

has a unique solution for any initial distribution G_0. A special case of the growth rate that we consider is the familiar replicator dynamics that was introduced by Taylor and Jonker (1978) for distributions with a finite support, and by Oechssler and Riedel (1999) for general distributions. In the case of the replicator dynamics, the distribution of types at time t, G_t, evolves according to the differential equation

$$\dot{G_t}(S) = \int_S \left[f(\tau^i, G_t) - f(G_t, G_t) \right] \mathrm{d}G_t(\tau^i), \qquad S \subseteq T, \tag{4.13}$$

where

$$f(\tau^i, G_t) \equiv \int_T f^i(\tau^i, \tau^j)\mathrm{d}G_t(\tau^j), \tag{4.14}$$

is the expected true equilibrium payoff of an individual of type τ^i at time t from an interaction with individual j drawn at random from the population, and

$$f(G_t, G_t) \equiv \int_T \int_T f^i(\tau^i, \tau^j) \mathrm{d}G_t(\tau^j) \mathrm{d}G_t(\tau^i), \tag{4.15}$$

is the expected true equilibrium payoffs when both individuals are drawn at random from the population at time t. That is, if the average performance of a subset of types $S \subseteq T$ is better than the average performance in the population, the relative weight in the population of the types in S increases, at the expense of other sets of types whose performance is below the average. The more general selection dynamics that we consider may be appropriate when the reproduction process of types is not purely biological, but rather relies on education or imitation (see e.g., Weibull 1995, Section 4.4).

Having defined the selection dynamics, we are now interested in the following question: starting from some initial distribution, G_0, how will the distribution of types, G_t, evolve over time with a regular, payoff-monotonic dynamics? To provide an answer, we first establish the following theorem. This theorem which is of independent interest, generalizes Theorem 1 in Samuelson and Zhang (1992) to the case of games with infinitely many strategies (Samuelson and Zhang, 1992, prove their result for the case of games with finitely many strategies). For preserving the coherence with our setting, we state the theorem for symmetric two-players games with a compact one-dimensional strategy space; the method of proof works however just as well for more general compact strategy spaces and for asymmetric games.

Theorem 1: *Let $T = [\underline{\tau}, \overline{\tau}] \subseteq R$ be a space of strategies, $f : T \times T \to R$ a continuous payoff function of a symmetric two-player game, and $g : T \times T \to R$ a regular, payoff monotonic growth-rate function. Let G_t be the population dynamics defined by the differential equation (4.12) with an initial distribution of strategies G_0 with support T. Suppose that $D \subseteq [\underline{\tau}, \overline{\tau}]$ is the subset of serially dominated strategies (those that do not survive the process of iterated elimination of strictly dominated strategies). Then the strategies in D are asymptotically eliminated from the population: Every iteratively dominated strategy $d \in D$ has an open neighborhood W_d for which $\lim_{t\to\infty} G_t(W_d) = 0$. In particular, if there is only one serially undominated strategy $u \in T \backslash D$, then G_t converges in distribution to the unit mass probability at u.*

Proof: *See the Appendix.*

We are now ready to state our main result:

Theorem 2: *Suppose that the payoffs in the pairwise interactions satisfy Properties 1-4. Then given any initial distribution of types with support T, the population of types will converge in distribution to a unit mass on some type $\widehat{\tau}$ under any regular, payoff-monotonic selection dynamics. The type $\widehat{\tau}$ is optimistic (i.e., $\widehat{\tau} > 0$) if Π^i_j and Π^i_{ij} have the same sign and pessimistic if Π^i_j and Π^i_{ij} have the opposite signs.*

Proof: From Lemma 2 we know that given properties 1-4, the artificial "types game" is strictly dominance solvable, with the solution being $(\widehat{\tau}, \widehat{\tau})$. Using Theorem 1 it therefore follows that under a regular, payoff monotonic growth-rate function, the population of types will converge in distribution to a unit mass on $\widehat{\tau}$. Finally, Lemma 1 ensures that $\widehat{\tau} > 0$ if Π^i_j and Π^i_{ij} have the same sign, and $\widehat{\tau} < 0$ if Π^i_j and Π^i_{ij} have the opposite signs. ■

5 Conclusion

We have shown how the pressures of explicit, dynamic evolutionary processes select for moderate optimism rather than for realism, when fitness is gained through interactions of either a competitive nature and strategic substitutes or cooperative nature and strategic complements. According to this insight, the well-documented phenomenon of overconfidence and unrealistic high self-esteem of individuals may be due to a bias that "pays" well in many kinds of *strategic* settings.

Clearly, the way humans evaluate their environment has evolved along the generations via conflicts with both natural hazards and strategic social interactions with other individuals or groups of individuals. The premises of our model are therefore far from being all-encompassing. And in practice, society is composed of heterogeneous individuals who may differ from one another in their degree of optimism/pessimism, unlike the long-run equilibrium in our model where all individuals share the same attitude. Thus, our modest aim was to point at one possible source for the optimism that is so frequently observed in

the process of decision making. Searching for competing and complementing evolutionary insights for this and similar behavioral puzzles is a challenge for future research.

6 Appendix

Proof of Proposition 2: Consider first the case where $b < 0$ (strategic complements). Let $A = [\underline{\tau}, \underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}] \cup [\overleftarrow{\tau} + \frac{\varepsilon}{2}, \overline{\tau}]$ be a compact subset of the set of types, $[\underline{\tau}, \overline{\tau}]$. Then $\{V_\tau : \tau \in A\}$, where V_τ is defined in (3.7), is an open cover of the compact set A, and therefore has a finite sub-cover $V_{\tau_1}, \ldots, V_{\tau_n}$. Since $\lim_{t\to\infty} G_t(V_{\tau_k}) = 0$ for $k = 1, \ldots, n$, there exists a time t_ε such that for $t > t_\varepsilon$ we have $G_t(V_{\tau_k}) < \frac{\varepsilon}{2nM}$, $k = 1, \ldots, n$, where $M = \max\{\overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right), \left(\underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}\right) - \underline{\tau}\}$. Hence, for $t > t_\varepsilon$ we have

$$G_t(A) \leq \sum_{k=1}^{n} G_t(V_{\tau_k}) < \frac{\varepsilon}{2M}.$$

Therefore, the average type in the population, θ, satisfies the following inequalities:

$$\begin{aligned} \theta &< \frac{\varepsilon}{2M}\overline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) = \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2M}\left(\overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right)\right) \qquad \text{(A.1)} \\ &\leq \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} = \overleftarrow{\tau} + \varepsilon, \end{aligned}$$

and

$$\begin{aligned} \theta &> \frac{\varepsilon}{2M}\underline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}\right) = \left(\underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2M}\left(\left(\underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}\right) - \underline{\tau}\right) \qquad \text{(A.2)} \\ &\geq \left(\underline{\tau}_{\rightarrow} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2} = \underline{\tau}_{\rightarrow} - \varepsilon. \end{aligned}$$

These inequalities imply that for every $\varepsilon > 0$, there exists a time t_ε such that for every $t > t_\varepsilon$, the average type θ is in the interval $[\underline{\tau}_{\rightarrow} - \varepsilon, \overleftarrow{\tau} + \varepsilon]$.

Next, note from equation (3.6) that the slope of the best-response function of type τ^i in the artificial types game in the (τ^i, τ^j) space is given by

$$-\frac{2b(4-b^2)^2(4-3b^2)}{(16-8b^2+b^4p)^2}.$$

Since $b < 0$, this expression is negative and less than 1 in absolute value. Hence, fixing the value of θ, there exists a unique symmetric Nash equilibrium in the artificial types game. Moreover, since $b < 0$, equation (3.6) shows that $BR^i(\cdot; p, \theta)$ is decreasing with θ. Hence,

the "highest" symmetric Nash equilibrium in the types game is attained when $\theta = \underrightarrow{\tau} - \varepsilon$ and the "lowest" equilibrium is attained when $\theta = \overleftarrow{\tau} + \varepsilon$. Let the highest and lowest symmetric Nash equilibria be $(\overleftarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon})$ and $(\underrightarrow{\tau_\varepsilon}, \underrightarrow{\tau_\varepsilon})$, respectively. That is, $\overleftarrow{\tau_\varepsilon}$ and $\underrightarrow{\tau_\varepsilon}$ are the solutions to the equations $\overleftarrow{\tau_\varepsilon} = BR^i(\overleftarrow{\tau_\varepsilon}; p, \underrightarrow{\tau} - \varepsilon)$ and $\underrightarrow{\tau_\varepsilon} = BR^i(\underrightarrow{\tau_\varepsilon}; p, \overleftarrow{\tau} + \varepsilon)$.

Noting from equation (3.5) that since $b < 0$, $\frac{\partial (f^i)^2(\tau^i, \tau^j; p, \theta)}{\partial \tau^i \partial \theta} = \frac{bp}{4} < 0$, it follows that if $\theta < \widetilde{\theta}$ and $\tau^i < \widetilde{\tau}^i$, then

$$f^i(\widetilde{\tau}^i, \tau^j; p, \theta) - f^i(\tau^i, \tau^j; p, \theta) \geq f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) - f^i(\tau^i, \tau^j; p, \widetilde{\theta}).$$

As a result, $f^i(\widetilde{\tau}^i, \tau^j; p, \theta) < f^i(\tau^i, \tau^j; p, \theta)$ implies $f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) < f^i(\tau^i, \tau^j; p, \widetilde{\theta})$ and $f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) > f^i(\tau^i, \tau^j; p, \widetilde{\theta})$ implies $f^i(\widetilde{\tau}^i, \tau^j; p, \theta) > f^i(\tau^i, \tau^j; p, \theta)$. These inequalities imply in turn that types above $\overleftarrow{\tau_\varepsilon}$ are serially dominated in the game played from t_ε onward, while types below $\underrightarrow{\tau_\varepsilon}$ are serially dominated from t_ε onward. By Theorem 1 below, this implies that types outside $[\underrightarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon}]$ get asymptotically extinct, implying (by the definition of $\overleftarrow{\tau}$ and $\underrightarrow{\tau}$) that $\overleftarrow{\tau} \leq \overleftarrow{\tau_\varepsilon}$ and $\underrightarrow{\tau} \geq \underrightarrow{\tau_\varepsilon}$ for every $\varepsilon > 0$. Since $\overleftarrow{\tau_\varepsilon}$ and $\underrightarrow{\tau_\varepsilon}$ are continuous functions of ε, we have for $\varepsilon = 0$ that

$$\begin{aligned} \overleftarrow{\tau} &\leq \inf_{\varepsilon > 0} \tau_\varepsilon \equiv \overleftarrow{\tau_0} = BR^i(\overleftarrow{\tau_0}; p, \underrightarrow{\tau}), \\ \underrightarrow{\tau} &\geq \sup_{\varepsilon > 0} \tau_\varepsilon \equiv \underrightarrow{\tau_0} = BR^i(\underrightarrow{\tau_0}; p, \overleftarrow{\tau}). \end{aligned}$$

Subtracting the second inequality from the first, using equation (3.6), and rearranging terms, yields:

$$0 \leq \overleftarrow{\tau} - \underrightarrow{\tau} \leq \overleftarrow{\tau_0} - \underrightarrow{\tau_0} = \frac{bp\,(2-b)\,(b+2)^2}{pb^4 + 4pb^3 + 4b^2 - 4pb^2 - 8b - 8bp - 16} \left(\overleftarrow{\tau} - \underrightarrow{\tau}\right). \qquad \text{(A.3)}$$

The coefficient of $\overleftarrow{\tau} - \underrightarrow{\tau}$ on the right side of the inequality is less than 1, implying that $\overleftarrow{\tau} = \underrightarrow{\tau} = \tau^*(p)$ as desired.

We now consider the case where $b > 0$ (strategic complements). Then, equation (3.5) implies that $\frac{\partial (f^i)^2(\tau^i, \tau^j; p, \theta)}{\partial \tau^i \partial \theta} = \frac{bp}{4} > 0$, so if $\theta < \widetilde{\theta}$ and $\tau^i < \widetilde{\tau}^i$, then

$$f^i(\widetilde{\tau}^i, \tau^j; p, \theta) - f^i(\tau^i, \tau^j; p, \theta) \leq f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) - f^i(\tau^i, \tau^j; p, \widetilde{\theta}).$$

As a result, $f^i(\widetilde{\tau}^i, \tau^j; p, \theta) > f^i(\tau^i, \tau^j; p, \theta)$ implies $f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) > f^i(\tau^i, \tau^j; p, \widetilde{\theta})$ and $f^i(\widetilde{\tau}^i, \tau^j; p, \widetilde{\theta}) < f^i(\tau^i, \tau^j; p, \widetilde{\theta})$ implies $f^i(\widetilde{\tau}^i, \tau^j; p, \theta) < f^i(\tau^i, \tau^j; p, \theta)$. Since $b > 0$, equation (3.6) implies

that $BR^i(\cdot; p, \theta)$ is upward sloping in the (τ^i, τ^j) space and moreover it is increasing with θ. Hence, the highest best-response of i intersects the lowest bets-response of j at $(\overleftarrow{\tau_\varepsilon}, \underrightarrow{\tau_\varepsilon})$. This implies in turn that types above $\overleftarrow{\tau_\varepsilon}$ for i and below $\underrightarrow{\tau_\varepsilon}$ for player j are serially dominated in the game played from t_ε onward. By Theorem 1 below, types outside $[\underrightarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon}]$ get asymptotically extinct, implying (by the definition of $\overleftarrow{\tau}$ and $\underrightarrow{\tau}$) that $\overleftarrow{\tau} \leq \overleftarrow{\tau_\varepsilon}$ and $\underrightarrow{\tau} \geq \underrightarrow{\tau_\varepsilon}$ for every $\varepsilon > 0$. Since $\overleftarrow{\tau_\varepsilon}$ and $\underrightarrow{\tau_\varepsilon}$ are continuous functions of ε, we have for $\varepsilon = 0$:

$$\begin{aligned} \overleftarrow{\tau} &\leq \inf_{\varepsilon>0} \tau_\varepsilon \equiv \overleftarrow{\tau_0} = BR^i(\underrightarrow{\tau_0}; p, \overleftarrow{\tau}), \\ \underrightarrow{\tau} &\geq \sup_{\varepsilon>0} \tau_\varepsilon \equiv \underrightarrow{\tau_0} = BR^i(\overleftarrow{\tau_0}; p, \underrightarrow{\tau}). \end{aligned}$$

Subtracting the second inequality from the first, using equation (3.5), and rearranging terms, yields:

$$\begin{aligned} \overleftarrow{\tau_0} - \underrightarrow{\tau_0} &= \frac{bp\,(2-b)\,(4b+8-b^3-2b^2)}{2\,(16-8b^2+pb^4)}\left(\overleftarrow{\tau} - \underrightarrow{\tau}\right) \\ &\quad + \frac{b\,(pb^4+4b^2-12pb^2+16p)}{2\,(16-8b^2+pb^4)}\left(\overleftarrow{\tau_0} - \underrightarrow{\tau_0}\right). \end{aligned}$$

This implies in turn that

$$0 \leq \overleftarrow{\tau} - \underrightarrow{\tau} \leq \overleftarrow{\tau_0} - \underrightarrow{\tau_0} = \frac{(-4+b^2)\,(2-b)\,bp}{pb^4-4b^3p+4b^2-4pb^2+8b+8bp-16}\left(\overleftarrow{\tau} - \underrightarrow{\tau}\right). \qquad \text{(A.4)}$$

Except for the extreme case where $b = 1$ and $p = 1$, the coefficient of $\overleftarrow{\tau} - \underrightarrow{\tau}$ is strictly smaller than 1, implying that $\overleftarrow{\tau} = \underrightarrow{\tau} = \tau^*(p)$ as desired. ■

Proof of Theorem 1: Let D_n be the set of strategies that do not survive n or less rounds of iterated elimination of strictly dominated strategies, so $D = \cup_{n=0}^{\infty} D_n$. Denote also by $U_n = T \setminus D_n$ the set of strategies that do survive n rounds of iterated elimination of strictly dominated strategies. We prove by induction on n that U_n is compact, and every eliminated strategy $d \in D_n$ has an open neighborhood W_d for which $\lim_{t\to\infty} G_t(W_d) = 0$.

Since $D_0 = \emptyset$ and $U_0 = T$, the claim holds for $n = 0$. If D_1 is empty as well, i.e. no strategies are strictly dominated, then the claim holds vacuously. So from now on assume that $D_1 \neq \emptyset$. Suppose, by induction, that the claim holds for $n < k$.

We first prove that U_k is compact. Indeed, let $d \in D_k$ be round-k dominated by the strategy $x \in T$, that is for every $y \in U_{k-1}$

$$f(x, y) - f(d, y) > 0.$$

Since f is continuous, $f(x,y)-f(d,y)$ is continuous in y, and therefore attains its minimum on U_{k-1}, as this set is compact by the induction hypothesis. Hence by (4.9)

$$\rho(x,d) \equiv \min_{y\in U_{k-1}} [f(x,y)-f(d,y)] > 0. \tag{A.5}$$

Furthermore, the function $\rho(x,d)$ is continuous since $[f(x,y)-f(d,y)]$ is. Therefore, for every $\varepsilon>0$, the set of strategies which are dominated by x by a payoff difference of at least ε, $\{d:\rho(x,d)>\varepsilon\}$, is open. Consequently, the set of strategies dominated up to round k

$$D_k = D_{k-1} \cup \bigcup_{x\in U_{k-1}} \bigcup_{\varepsilon>0} \{d:\rho(x,d)>\varepsilon\},$$

is open as a union of open sets, and $U_k = T \setminus D_k$ is compact, as required.

We now turn to complete the inductive step, and prove that every eliminated strategy $d\in D_k$ has an open neighborhood W_d for which $\lim_{t\to\infty} G_t(W_d)=0$. Indeed, let d be dominated by x in one of the rounds up to k. Then if $k>1$, not only does x perform better than d against strategies in U_{k-1}, but it also does so against some strategies in an open subset of D_{k-1}: Since f is continuous, the set

$$B = \{y\in T : f(x,y)-f(d,y)\le 0\},$$

is a compact subset of the open set D_{k-1}. Hence B is a *proper* subset of D_{k-1}, as D_{k-1} is open by the induction hypothesis (except when $k=1$, in which case $D_{k-1}=\emptyset$, and $B=\emptyset$).

This implies that for some positive number s, we have

$$\lim_{t\to\infty} G_t(C) = 0,$$

where

$$C = \{y\in D_{k-1} : f(x,y)-f(d,y)\le s\}.$$

Indeed, for $k=1$, $D_{k-1}=D_0=\emptyset$, so for whatever positive s chosen we have $G_t(C)=0$, so $\lim_{t\to\infty} G_t(C)=0$. For $k>1$, let

$$s = \frac{1}{2}\sup_{y\in D_{k-1}} [f(x,y)-f(d,y)], \tag{A.6}$$

which is positive since as explained above, $B \subsetneq D_{k-1}$. With this s, the set C is a closed subset of D_{k-1} (because $f(x,y) - f(d,y)$ is continuous in y), and hence compact. By the induction hypothesis, every $y \in D_{k-1}$ has an open neighborhood $W_y \ni y$ such that $\lim_{t\to\infty} G_t(W_y) = 0$. Since $C \subseteq \bigcup_{y\in C} W_y$, the compactness of C implies that there exist $y_1, \dots y_m \in C$ such that $C \subseteq \bigcup_{i=1}^m W_{y_i}$. Therefore, $G_t(C) \leq \sum_{i=1}^m G_t(W_{y_i})$ and hence $\lim_{t\to\infty} G_t(C) = 0$, as claimed.

Observe now the following considerations. The continuity of f and ρ imply that there are open neighborhoods $V_x \ni x$ and $W_d \ni d$ such that for every $x' \in \overline{V}_x$, $d' \in \overline{W}_d$

$$\inf_{y\in D_{k-1}\setminus C} [f(x',y) - f(d',y)] \geq \frac{s}{2} > 0, \tag{A.7}$$

and

$$\min_{y\in U_{k-1}} [f(x',y) - f(d',y)] = \rho(x',d') \geq \frac{\rho(x,d)}{2} > 0. \tag{A.8}$$

Thus, against any strategy $y \notin C$, every strategy $x' \in \overline{V}_x$ outperforms every $d' \in \overline{W}_d$ by at least

$$\varepsilon = \min\left\{\frac{s}{2}, \frac{\rho(x,d)}{2}, \frac{1}{2}\right\} > 0, \tag{A.9}$$

i.e.,

$$\inf_{y\in T\setminus C} [f(x',y) - f(d',y)] \geq \varepsilon. \tag{A.10}$$

At the same time, since f is continuous on the compact domain T , there exists a bound M such that $|f| \leq M$; and by (A.6), there exists a time $\overline{t}$ such that for $t \geq \overline{t}$ we have $G_t(C) < \frac{\varepsilon}{8M}$ and $G_t(T \setminus C) > 1 - \varepsilon$. Altogether this implies that for $x' \in \overline{V}_x, d' \in \overline{W}_d$ and $t \geq \overline{t}$

$$\begin{gathered} f(x',G_t) - f(d',G_t) = \int_T [f(x',\cdot) - f(d',\cdot)]\, dG_t = \\ \int_C [f(x',\cdot) - f(d',\cdot)]\, dG_t + \int_{T\setminus C} [f(x',\cdot) - f(d',\cdot)]\, dG_t > \\ (-2M)\,\frac{\varepsilon}{8M} + \varepsilon(1-\varepsilon) \geq -\frac{\varepsilon}{4} + \varepsilon(1-\frac{1}{2}) = \frac{\varepsilon}{4}. \end{gathered} \tag{A.11}$$

By the continuity of f, (A.11) holds also when G_t is replaced by any probability measure

$\mu \in A \equiv \overline{\{G_t\}_{t\geq \overline{t}}}$, the closure of $\{G_t\}_{t\geq \overline{t}}$ in the topology of weak convergence of probability measures.

Now, by the payoff monotonicity of the growth-rate function g, for every $\mu \in A$, $x' \in \overline{V}_x$ and $d' \in \overline{W}_d$

$$g(x', \mu) - g(d', \mu) > 0.$$

The continuous function $[g(x', \mu) - g(d', \mu)]$ attains its minimum on the compact set $\overline{V}_x \times \overline{W}_d \times A$. Therefore, there is in fact $\delta > 0$ for which

$$g(x', G_t) - g(d', G_t) \geq \delta, \qquad \text{for } \; x' \in \overline{V}_x, \;\; d' \in \overline{W}_d, \;\text{ and }\; t \geq \overline{t}. \tag{A.12}$$

A fortiori, (A.12) holds also if we replace $g(x', G_t)$ and $g(d', G_t)$ by their averages in $\overline{V}_x$ and $\overline{W}_d$, respectively. Thus for $t \geq \overline{t}$

$$\frac{\int_{\overline{V}_x} g(\cdot, G_t)\mathrm{d}G_t}{G_t(\overline{V}_x)} - \frac{\int_{\overline{W}_d} g(\cdot, G_t)\mathrm{d}G_t}{G_t(\overline{W}_d)} \geq \delta. \tag{A.13}$$

Hence, by (4.12), for $t \geq \overline{t}$,

$$\frac{G_t(\overline{V}_x)}{G_t(\overline{W}_d)} \geq \frac{G_{\overline{t}}(\overline{V}_x)}{G_{\overline{t}}(\overline{W}_d)} \exp[\delta(t - \overline{t})] \rightarrow_{t\to\infty} \infty. \tag{A.14}$$

Therefore, $\lim_{t\to\infty} G_t(W_d) = 0$, as required. ■

7 References

Babcock L. and G. Loewenstein (1997), "Explaining Bargaining Impasse: The Role of Self-Serving Biases," *Journal of Economic Perspectives,"* **11(1)**, pp. 109-126.

Benabou R. and J. Tirole (1999a), "Self-Confidence: Intrapersonal Strategies," mimeo, Princeton University.

Benabou R. and J. Tirole (1999b), "Self-Confidence and Social Interactions," mimeo, Princeton University.

Benos A.V. (1998), "Aggressiveness and Survival of Overconfident Traders," *Journal of Financial Markets* **1**, pp. 353-383.

Bergman N. and Y. Bergman (2000), "Ecologies of Preferences with Envy as an Antidote to Risk-Aversion in Bargaining", mimeo, The Hebrew University of Jerusalem.

Bester H. and W. Güth (1998), "Is Altruism Evolutionary Stable?", *Journal of Economic Behavior and Organization* **34(2),** pp. 211-221.

Bolle F. (2000), "Is Altruism Evolutionarily Stable? And Envy and Malevolence? - Remarks on Bester and Güth," *Journal of Economic Behavior and Organization* **42(1),** pp. 131-133.

Brander J. and T. Lewis (1986), "Oligopoly and Financial Structure: The Limited Liability Effect," *American Economic Review* **76**, pp. 956- 970.

Brander J. and B.J. Spencer (1985), "Export Subsidies and International Market Share Rivalry," *Journal of International Economics* **18**, pp. 83-100.

Bulow, J., J. Geanakoplos, and P. Klemperer (1985), "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*; **93(3)**, pp. 488-511.

Brocas I. and J. Carrillo (1999), "Entry Mistakes, Entrepreneurial Boldness and Optimism," CEPR D.P. 2213.

Camerer C. and D. Lovallo (1999), "Overconfidence and Excess Entry: An Experimental Approach," *American Economic Review* **89(1)**, pp. 306-318.

Cooper A.C., W. Dunkelberg, and C. Woo.(1988), "Entrepreneur's percieved Chances of Success," *Journal of Business Venturing*, **3**, pp. 97-108.

Dekel E., J. Ely, and O. Yilankaya (1998), "Evolution of Preferences," mimeo, Northwestern University

Dekel E. and S. Scotchmer (1999), "On the Evolution of Attitudes toward Risk in Winner-Take-All Games," *Journal of Economic Theory* **87**, pp. 95-124.

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**, pp. 281-295.

Eaton J. and G.M. Grossman (1986), "Optimal Trade and Industrial Policy under Oligopoly," *Quarterly Journal of Economics*, pp. 383-406.

Fershtman C. and K. Judd (1987), "Incentive Equilibrium in Oligopoly," *American Economic Review* **77(5)**, pp. 927-940

Fershtman C., K. Judd and E. Kalai (1991), "Observable Contracts: Strategic Delegation and Cooperation," *International Economic Review* **32(3)**, pp. 551-559.

Fershtman C. and E. Kalai (1997), "Unobserved Delegation," *International Economic Review* **38(4)**, pp. 763-774.

Fershtman C. and Y. Weiss (1997), "Why do We Care about what Others Think about Us?," in: Ben Ner, A. and L. Putterman (eds.), *Economics, Values and Organization*, Cambridge University Press, Cambridge MA.

Fershtman C. and Y. Weiss (1998), "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics* **70**, pp. 53-74.

Fleming J. and J.M. Darley (1986), "Perceiving Intension in Constrained Behavior: The Role of Purposeful and Constrained Action Cues in Correspondence Bias Effects," mimeo, Princeton University.

Frank R.H. (1987), "If Homo Economicus Could Choose His Own Utility Function, Would He Choose One With a Conscience?" *American Economic Review* **77(4)**, pp. 593-604.

Frank R.H. (1988), *Passions Within Reason – The Strategic Role of the Emotions,* W.W. Norton & Company, New York.

Güth W. and M. Yaari (1992), "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in Witt, U. (ed.), *Explaining Forces and Changes: Approaches to Evolutionary Economics,* University of Michigan Press.

Green J. (1992), "Commitment with Third Parties," *Annales d'Economie et de Statistique*, **25- 26**, 101- 121.

Heifetz A. and E. Segev (2001), "The Evolutionary Role of Toughness in Bargaining," Foerder DP #25-01, available at http://econ.tau.ac.il/papers/foerder/25-2001.pdf

Huck S. and J. Oechssler (1998), "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior* **28**, pp. 13-24.

Huck S., G. Kirchsteiger, and J. Oechssler (1997), "Learning To Like What You Have – Explaining the Endowment Effect," mimeo, Humboldt University, Berlin.

Katz M. (1991), "Game-Playing Agents: Unobservable Contracts as Precommitments," *Rand Journal of Economics* **22**, pp. 307- 328.

Koçkesen L., E.A.. Ok, and R. Sethi (2000a), "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior*, **31**, pp. 303-310.

Koçkesen L., E.A. Ok, and R. Sethi (2000b), "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory,* **92**, pp. 274-299.

Kyle A.S. and A. Wang (1997), "Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?" *The Journal of Finance* **LII**, pp. 2073-2090.

Lazear E. and S. Rosen (1981), "Rank-Order Tournaments as Optimum Labor Contracts," *The Journal of Political Economy*, **89(5)**, pp. 841-864.

Lehman D. R., and R.E. Nisbett (1985). "Effects of Higher Education on Inductive Reasoning," Unpublished manuscript, University of Michigan (cited in D. Myers, 1998, *Psychology,* Fifth Edition, Worth Publishers Inc.).

Lewinsohn P.M., W. Mischel, W. Chaplin, and R. Barton (1980), "Social Competence and Depression: The Role of Illusory Self-Perceptions," *Journal of Abnormal Psychology* **898**, pp. 203-212.

Moulin H. (1984), "Dominance Solvability and Cournot Stability," *Mathematical Social Sciences,"* **7(1)**, pp. 83-102.

Myers D. (1998), *Psychology,* Fifth Edition, Worth Publishers Inc.

Oechssler J. and F. Riedel (2001), "Evolutionary Dynamics on Infinite Strategy Spaces," *Economic Theory, **17**,* pp. 141-162..

Ok E.A. and F. Vega-Redondo (2001), "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, **97**, pp. 231-254.

Possajennikov A. (2000), "On The Evolutionary Stability of Altruistic and Spiteful Prefer-

ences," *Journal of Economic Behavior and Organization* **42(1)** pp. 125-129 (preliminary version appeared as Possajenikov (1999), CentER working paper 9956, Tilburg University).

Robson A.R. (1996a), "A Biological Basis for Expected and Non-Expected Utility," *Journal of Economic Theory* **68**, pp. 397-424.

Robson A.R. (1996b), "The Evolution of Attitudes to Risk: Lottery Tickets and Relative Wealth," *Games and Economic Behavior* **14**, pp. 190-207.

Rogers A.R. (1994), "Evolution of Time Preference by Natural Selection," *American Economic Review* **84**, pp. 460-481.

Rogoff K. (1985), "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, **100(4)**, pp. 1169-1189.

Samuelson L. and J. Zhang (1992), "Evolutionary Stability in Asymmetric Games," *Journal of Economic Theory* **57**, pp. 363-391.

Sandolm W. (2001), ""Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Econmic Dynamics* **4**, pp. 637-679.

Sandroni A. (2000), "Do Markets Favor Agents Able to Make Accurate Predictions," *Econometrica,* **68**, pp. 1303-1341.

Schelling T., (1960), *The Strategy of Conflict*, Cambridge MA: Harvard University Press.

Seligman M. E. P. and , P. Schulman (1986), "Explanatory Style as a Predictor of Productivity and Quitting Among Life Insurance Sales Agents," *Journal of Personality and Social Psychology*, **50**, pp. 832-838.

Svenson O. (1981), "Are We All Less Risky Drivers and More Skillful than our Fellow Drivers?" *Acta Psychologica* **47(2)**, pp. 143-148.

Taylor S.E. and J.D. Brown (1988), "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin* **103**, pp. 193-210.

Taylor P. and L. Jonker (1978), "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences* **40**, pp. 145-156.

Vega-Redondo F. (1997), "The Evolution of Walrasian Behavior," *Econometrica* **65**, pp. 375-384.

Waldman M. (1994), "Systematic Errors and the Theory of Natural selection," *American Economic Review* **84(3)**, pp. 482-497.

Wason P. C. (1981), "The importance of cognitive illusions," *The Behavioral and Brain Sciences* **4**, p. 356.

Weibull J. (1995), *Evolutionary Game Theory*, MIT Press, Cambridge MA..