Economic Theory (2007) 32: 251–286
DOI 10.1007/s00199-006-0121-7

RESEARCH ARTICLE

Aviad Heifetz · Chris Shannon · Yossi Spiegel

The dynamic evolution of preferences

Received: 22 November 2004 / Accepted: 4 February 2006 / Published online: 7 June 2006

Abstract This paper develops a general methodology for characterizing the dynamic evolution of preferences in a wide class of strategic interactions. We give simple conditions characterizing the limiting distribution of preferences in general games, and apply our results to study the evolutionary emergence of overconfidence and interdependent preferences. We also show that this methodology can be adapted to cases where preferences are only imperfectly observed.

Keywords Evolution of preferences · Evolutionary stability · Overconfidence · Interdependent preferences

JEL Classification Numbers C72 · D01

Earlier drafts of this paper were circulated under the title: "The Evolution of Perception Biases" and some of these results appeared in the earlier working paper Heifetz, Shannon and Spiegel (2003). We thank three anonymous referees for their helpful comments. Shannon thanks the NSF for research support under grant SES-0351346.

A. Heifetz
The Economics and Management Department,
The Open University of Israel, Tel-Aviv, Israel

C. Shannon (✉)
Department of Economics, University of California,
549 Evans, Hall, Berkeley, CA 94709, USA
E-mail: cshannon@econ.berkeley.edu

Y. Spiegel
Recanati Graduate School of Business Administration,
Tel Aviv University, Tel-Aviv, Israel

1 Introduction

One of the cornerstones of economic analysis is the assumption that economic agents are rational and self-interested. Among other things, rationality typically includes the assumption that agents correctly perceive material payoffs and choose actions to maximize these payoffs. This assumption is often justified either formally or informally by appealing to evolutionary arguments. For example, in their classic work, Alchian (1950) and Friedman (1953) argue that profit maximization is a reasonable assumption for characterizing outcomes in competitive markets because firms that fail to maximize their profits will be driven out of the market by more profitable rivals. Similar arguments that consumers behave "as if" maximizing their material payoffs due to myriad market forces that exploit non-optimal behavior are pervasive.

Despite its central role in economic analysis, the assumption that economic agents behave as if maximizing their material payoffs is at odds with a large body of evidence from psychology and from experimental economics suggesting that individuals often pursue objectives other than actual payoff maximization. Observed departures from material payoff maximizing behavior routinely arise through actions that are altruistic or spiteful, that favor fairness or reciprocity, or that show concern for relative payoffs. Individuals often demonstrate inaccurate assessments of their environments, including their own abilities, payoffs, and probabilistic assessments. For example, people are frequently overconfident, with voluminous evidence suggesting that the ranks of the overconfident include judges (Guthrie et al. 2001), psychologists (Oskamp 1965), physicians (Christensen-Szalanski and Bushyhead 1981), engineers (Kidd 1970), entrepreneurs (Camerer and Lovallo 1999; Cooper et al. 1988), negotiators (Babcock and Loewenstein 1997), securities analysts (Froot and Frankel 1989; De Bondt and Thaler 1985), and managers (Russo and Schoemaker 1992). In many experiments, subjects sacrifice personal payoffs to reward or punish others, or for purely altruistic reasons. Evidence also suggests that the nature of these departures may vary with the context or nature of interaction.[1]

Motivated in part by this evidence, a large body of literature has recently emerged that examines the evolutionary foundations of preferences. Far from validating informal evolutionary arguments for payoff-maximizing behavior, this work shows that many different systematic departures from payoff maximization may survive evolutionary pressures in various models. Individuals in these models may have preferences that differ from their true material payoffs due, for example, to concerns about fairness (Güth and Yaari 1992, Huck and Oechssler 1998), social status (Fershtman and Weiss 1997, 1998), altruism (Bester and Güth 1998), spite (Possajennikov 2000, Bolle 2000), envy (Bergman and Bergman 2000), relative rather than absolute success (Koçkesen et al. 2000a, b), or overconfidence (Kyle and Wang 1997, Benos 1998). The main results in these papers show that such biases or dispositions may be evolutionarily stable in particular models, and thus immune to the appearance of rational "mutants" who maximize their actual material payoffs.[2]

[1] For example, see Chapter 2 of Camerer (2003) for an extensive discussion of this literature.

[2] The indirect evolutionary approach, where players' preferences rather than strategies evolve over time, is also employed by Dekel and Scotchmer (1999); Dufwenberg and Güth (1999);

In this paper, we extend these results in several directions. First, rather than focusing on a particular model or parametric specification for payoff and bias functions, we seek general conditions under which results about the evolution of preferences are valid. Second, we focus on characterizing the evolutionary dynamics of the distribution of preferences, and on characterizing the support of the limiting distribution, rather than on the evolutionary stability of certain population profiles as in other recent work on the evolution of preferences in general games, including Dekel et al. (2005) and Ely and Yilankaya (2001).[3] Developing a truly dynamic evolutionary model of the evolution of preferences is necessary for actually making predictions based on such models. Results based solely on evolutionary stability are essentially static in nature. As such, they can help explain the immunity of a particular population profile to sporadic mutations, but say nothing about whether competitive selection could ever lead to such population profiles given arbitrary initial states, and consequently say little about which distributions are actually likely to emerge. Thus it is important to establish results that are dynamic and that predict long-run outcomes corresponding to a wide range of initial population distributions.

The idea that in strategic situations players may gain an advantage from having an objective function different from actual payoff maximization dates back at least to Schelling (1960), and his discussion of the commitment value of decision rules. The same ideas can be seen in the work on strategic delegation (e.g., Fershtman and Judd 1987 and Katz 1991). Considered in this light, the intuition for the survival of various biases is fairly straightforward. Having a disposition has two effects on a player's payoff: a direct effect, through the player's own actions, and an indirect effect, by influencing other players' actions. The direct effect of a small nonzero degree of bias must always be negligible, as it results from a slight deviation from payoff-optimizing behavior, yet the indirect effect resulting from the induced changes in opponents' actions may easily not be negligible. In a companion paper (Heifetz, Shannon and Spiegel 2006), we demonstrate this general principle behind much of the work on the evolution of preferences, and show that the emergence of dispositions is in fact generic: in almost every game and for almost every family of distortions of a player's actual payoffs, some degree of this distortion is beneficial to the player. Consequently, any such distortions will not be driven out by any payoff-monotonic selection dynamics. In the current paper we complement these results by developing a methodology that allows us to uniquely characterize the limiting distribution of preferences in the population for a wide class of strategic interactions.

To establish our results we adopt the indirect evolutionary approach, positing that evolutionary selection dynamics operate on preferences based on the equilibrium payoffs that individuals with these preferences obtain. The individuals choose actions in the underlying game based on their perceived payoff functions, and then receive payoffs according to their actual payoff functions. Thus the evolutionary

Rogers (1994); Robson (1996a, b); Waldman (1994); and Vega-Redondo (1997). See also further references below.

[3] Other fully dynamic models of the evolution of preferences include Huck et al. (2006) who consider the emergence of an endowment effect in bargaining and Sandholm (2001) who considers individual dispositions towards particular strategies. Apart from the general context, these papers differ from ours in that the dynamics in Huck et al. (2006) is not shown to converge in the long-run, whereas Sandholm only studies 2×2 normal form games.

fitness of a particular type is based on the equilibrium payoffs in an artificial second stage "types" game with payoff functions given by the actual payoffs induced in equilibrium given the play of each type. We show that if this types game is dominance solvable, then given any initial distribution with full support and any payoff monotonic selection dynamics, the population distribution converges to a point mass at the unique type profile that survives the iterated elimination of strictly dominated strategies in the types game. Our results follow from the more general result that any serially dominated strategy must eventually become extinct under payoff-monotonic selection dynamics, which we establish by extending results of Samuelson and Zhang (1992) to games with a continuum of actions. This methodology allows us to determine the long run population dynamics in a relatively straightforward way.

The paper is organized as follows. Section 2 develops the evolutionary framework and contains our main results. In section 3 we illustrate applications of our methodology by studying two types of dispositions: perception biases (such as optimism and pessimism) and interdependent preferences (such as altruism and spite). In each case we are able to characterize the unique level of disposition to which the population converges over time under any payoff-monotonic selection dynamics. In section 4 we show that our methodology can be adapted to cases where preferences are only imperfectly observed. We consider two alternative settings: (i) preferences are perfectly observed in some fraction of interactions but unobserved in others, and (ii) the model involves costly signaling of preferences. In both settings we are able to completely characterize the limiting distribution of types. Section 5 extends our main example to cases in which there may be payoff uncertainty. Section 6 concludes, and the Appendix collects several lengthy proofs.

2 A general analysis

2.1 Payoffs and dispositions

Two players, i and j, engage in strategic interaction. The strategy spaces of the two players are $X^i = X^j \subset \mathbf{R}$. Typical strategies are denoted x^i and x^j. We consider a symmetric game in which the payoffs of the two players are given by

$$\Pi^i\left(x^i, x^j\right) \equiv \Pi\left(x^i, x^j\right), \quad \Pi^j\left(x^i, x^j\right) \equiv \Pi\left(x^j, x^i\right),$$

for a function $\Pi : \mathbf{R}^2 \to \mathbf{R}$.

In the course of their strategic interaction, the players perceive their payoffs to be

$$\begin{aligned} U^i\left(x^i, x^j, \tau^i\right) &\equiv \Pi^i\left(x^i, x^j\right) + B^i\left(x^i, x^j, \tau^i\right), \\ U^j\left(x^i, x^j, \tau^j\right) &\equiv \Pi^j\left(x^i, x^j\right) + B^j\left(x^i, x^j, \tau^j\right), \end{aligned} \tag{1}$$

where

$$B^i\left(x^i, x^j, \tau^i\right) \equiv B\left(x^i, x^j, \tau^i\right), \quad B^j\left(x^i, x^j, \tau^j\right) \equiv B\left(x^j, x^i, \tau^j\right),$$

for a function $B : \mathbf{R}^3 \to \mathbf{R}$. The functions B^i and B^j are the dispositions of players i and j, and τ^i and τ^j are the players' types.[4] Types are drawn from the compact set $T = [\underline{\tau}, \overline{\tau}] \subseteq \mathbf{R}$, where $\underline{\tau} < 0 < \overline{\tau}$. The dispositions drive a wedge between the objectives of the players, which are to maximize their perceived payoffs U^i and U^j, and their eventual realized payoffs Π^i and Π^j. In the next section we provide two examples for individual dispositions: perception biases (such as optimism or pessimism) and interdependent preferences (such as altruism or spite).

Moreover, as a normalization we assume that

$$B^i\left(x^i, x^j, 0\right) \equiv B^j\left(x^i, x^j, 0\right) \equiv 0.$$

That is, a type 0 player has no disposition and hence maximizes his actual payoff.[5]

Let $\Gamma = (X^i, X^j, \Pi^i, \Pi^j, B^i, B^j)$ denote the game in which players i and j choose actions x^i and x^j, respectively, to maximize their perceived payoffs, $U^i(\cdot, \tau^i)$ and $U^j(\cdot, \tau^j)$, and obtain true payoffs Π^i and Π^j. We will maintain the following assumption about Γ:

Assumption A The game Γ has a unique pure strategy equilibrium $(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j))$ for each $(\tau^i, \tau^j) \in T \times T$.

Because we adopt the indirect evolutionary approach, we will measure the "fitness" of a type through the payoffs that type achieves in equilibrium play of the underlying game Γ. Obtaining positive results about the convergence of types requires determining how these equilibrium payoffs change as the type profile (τ^i, τ^j) changes. For some classes of games it may be possible to characterize these changes in equilibrium payoffs uniformly across multiple equilibria. For such games, it may be possible to extend our basic methodology. For many games with multiple equilibria, however, the comparative statics of equilibrium payoffs in response to changes in the types τ^i and τ^j will differ for different equilibria. Deriving positive results concerning the convergence of evolutionary dynamics for such games would then require some theory of equilibrium selection, or a more explicit dynamic model in which the selection becomes endogenous. Because we seek general results that may hold across a variety of problems, we therefore maintain the assumption that the underlying game has a unique pure strategy equilibrium for each profile of types.

The issues involving multiple equilibria in our setting also highlight the differences between the dynamic results we establish and the static results regarding evolutionary stability that are the focus of much of the literature on the evolution of preferences. Because evolutionary stability is both local and static, characterizations of evolutionary stability can be derived in the presence of multiple equilibria in the underlying game by conditioning on a particular equilibrium. Moreover, most of these results seek characterizations of evolutionarily stable preference and

[4] We assume symmetry of payoffs and dispositions, because in what follows we restrict attention to a scenario in which the two players are drawn from one large population.

[5] Notice that this formulation in terms of an additive disposition term is equivalent to specifying instead that a player has preferences given by a utility function $U^i(x^i, x^j, \tau)$ such that $U^i(x^i, x^j, 0) \equiv \Pi^i(x^i, x^j)$. To see this, given such a utility function simply set $B^i(x^i, x^j, \tau) \equiv U^i(x^i, x^j, \tau) - \Pi^i(x^i, x^j)$.

action profiles without simultaneously addressing the issue of the existence of such profiles; not surprisingly, assumptions for such results may be more general.[6]

Simple well-known sufficient conditions can be given on the payoff and disposition functions under which Assumption A holds. For these conditions, we assume that Π^i, Π^j, B^i and B^j are C^2. In what follows, to simplify notation we often denote partial derivatives by

$$\Pi^i_i\left(x^i,x^j\right)\equiv\frac{\partial\Pi^i}{\partial x^i}\left(x^i,x^j\right),\qquad \Pi^i_j\left(x^i,x^j\right)\equiv\frac{\partial\Pi^i}{\partial x^j}\left(x^i,x^j\right),$$
$$\Pi^i_{ii}\left(x^i,x^j\right)\equiv\frac{\partial^2\Pi^i}{(\partial x^i)^2}\left(x^i,x^j\right),\qquad \Pi^i_{ij}\left(x^i,x^j\right)\equiv\frac{\partial^2\Pi^i}{\partial x^i\partial x^j}\left(x^i,x^j\right),$$

with analogous notation for partial derivatives of Π^j, B^i, B^j, U^i, and U^j.

We note several simple sufficient conditions under which Assumption A holds below.

Assumption A1 U^i is C^2 and differentiably strictly concave in x^i, i.e., $U^i_{ii}(x^i,x^j,\tau^i)<0$ for all $(x^i,x^j,\tau^i)\in\mathbf{R}^2\times T$; analogously for U^j.

Assumption A2 U^i is C^2 and there exists $\varepsilon>0$ such that $|U^i_{ij}(x^i,x^j,\tau^i)|<(1-\varepsilon)|U^i_{ii}(x^i,x^j,\tau^i)|$ for all $(x^i,x^j,\tau^i)\in\mathbf{R}^2\times T$; analogously for U^j.

Assumption A1 ensures that players' best response functions are implicitly defined by the first order conditions

$$U^i_i\left(x^i,x^j,\tau^i\right)=\Pi^i_i\left(x^i,x^j\right)+B^i_i\left(x^i,x^j,\tau^i\right)=0,$$
$$U^j_j\left(x^i,x^j,\tau^j\right)=\Pi^j_j\left(x^i,x^j\right)+B^j_j\left(x^i,x^j,\tau^j\right)=0,$$

provided their best responses are always interior. When the strategy sets X^i and X^j are convex and best responses are well-defined, Assumption A2 ensures that the slope of each player's best-response function is uniformly less than 1 in absolute value. This implies in turn that there is a unique Nash equilibrium in this game.[7] Assumption A2 also ensures the stronger result that myopic best-reply dynamics, in which each player plays a best reply to the previous action of the opponent, will converge to the unique Nash equilibrium. This may justify the assumption that individuals play the unique Nash equilibrium even if initially they do not observe each other's type.

Under Assumptions A1 and A2, the unique interior Nash equilibrium in the game given types (τ^i,τ^j), $(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j))$, is implicitly defined by the following two equations:

$$\Pi^i_i\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j)\right)+B^i_i\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j),\tau^i\right)=0,$$
$$\Pi^j_j\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j)\right)+B^i_j\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j),\tau^j\right)=0.$$

[6] For example, existence is an issue for Dekel et al. (2005) in the case of partially observable preferences as a by-product of the tension between efficiency, which they show is necessary for stability in the case of observable preferences, and Nash equilibrium, which is necessary for stability in the unobservable case.

[7] The uniformity requirement is needed to guarantee that the best-response functions actually intersect, and thus that an equilibrium exists.

1.3 The types game

To study the evolution of dispositions, we adopt the indirect evolutionary approach. Given Assumption A, the true payoffs of players i and j in the unique Nash equilibrium of the primitive game Γ are:

$$f^i\left(\tau^i, \tau^j\right) \equiv \Pi^i\left(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)\right) \tag{2}$$
$$f^j\left(\tau^i, \tau^j\right) \equiv \Pi^j\left(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)\right).$$

We now consider a "types game" in which each player chooses a type $\tau^i, \tau^j \in T = [\underline{\tau}, \overline{\tau}]$, and receives a payoff according to the fitness functions f^i and f^j. Notice that since the original underlying game is symmetric, and all individuals have the same true payoff function Π, the types game is also symmetric, that is, $f^i(\tau^i, \tau^j) = f^j(\tau^j, \tau^i)$ for each $(\tau^i, \tau^j) \in T \times T$. In what follows, we will therefore omit the superscripts i and j from the function $f(\cdot, \cdot)$ whenever this does not cause confusion.

To characterize the dynamic evolution of types, we will assume for many of our results that this derived types game is dominance solvable.

Assumption B The types game $\mathcal{T} = (T, T, f^i, f^j)$ is dominance solvable.

As with Assumption A, the strength of Assumption B is a function of the strength of the results we seek. We show both that the distribution of types in the population converges under a broad class of selection dynamics, and characterize the support of the limit distribution. As with Assumption A, the particular features of certain classes of games may allow our general methodology to be used to obtain these results in the absence of dominance solvability. A notable example is given below in Theorem 3, in which we characterize the limit distribution for fitness games with strategic complementarities.

We note that as with Assumption A, well-known conditions on the fitness functions f^i and f^j are sufficient to ensure that Assumption B holds.

Assumption B1 f^i is C^2 and differentiably strictly concave in τ^i, i.e., $f^i_{ii}(\tau^i, \tau^j) < 0$ for all $(\tau^i, \tau^j) \in T \times T$; analogously for f^j.

Assumption B2 $|f^i_{ij}(\tau^i, \tau^j)| < |f^i_{ii}(\tau^i, \tau^j)|$ for all $(\tau^i, \tau^j) \in T \times T$; analogously for f^j.

Assumption B1 ensures that the types game is well-behaved. Assumption B2 implies that the slope of each player's best-response function in the types game is less than 1 in absolute value.[8] Under Assumption B1, an interior Nash equilibrium in the types game is defined implicitly by the following pair of equations:

$$f^i_i\left(\tau^i, \tau^j\right) = 0, \quad f^j_j\left(\tau^i, \tau^j\right) = 0.$$

[8] We have chosen to state Assumptions B1 and B2 in terms of the derived payoff functions in the type game. There are corresponding conditions in terms of the payoff functions in the original underlying game, Π, but they involve complex expressions that include third-order derivatives of Π which do not yield much additional insight. As we show in the applications that follow, we suspect it will be easier to simply check these conditions directly on the derived fitness functions.

Since the types game is symmetric, the unique Nash equilibrium in the types game is also symmetric and given by $(\widehat{\tau}, \widehat{\tau})$. If $\widehat{\tau}$ is interior it is defined implicitly by the equation $f_i^i(\widehat{\tau}, \widehat{\tau}) \equiv 0$.

Lemma 1 *Under Assumption A and Assumptions B1 and B2, the types game $\mathcal{T}$ is dominance solvable. The unique strategy profile that survives iterated elimination of strictly dominated strategies is the unique symmetric Nash equilibrium $(\widehat{\tau}, \widehat{\tau})$.*

Moreover, if $(\widehat{x}^i(0,0), \widehat{x}^j(0,0)) \in int\,(X^i \times X^j)$, then under Assumptions A1–A2 and B1–B2,

$$\begin{aligned} sign\,\widehat{\tau} &= sign\, f_i^i(0,0) \\ &= sign\,\Pi_j^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))\Pi_{ij}^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))B_{i\tau^i}^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0), 0). \end{aligned}$$

Proof To prove dominance solvability of the types game, we appeal to Theorem 4 in Moulin (1984) (see also Rosen 1965). By this result, it suffices to show that

(i) The strategy set of each player is a one-dimensional compact interval;
(ii) The payoff function of each player is continuous, twice differentiable, and strictly concave with respect to the player's strategy;
(iii) The slope of each player's best-response function is less than 1 in absolute value.[9]

Condition (i) is satisfied in the types game because the set of possible types for each player is the compact interval $T = [\underline{\tau}, \overline{\tau}]$. Assumptions B1 and B2 ensure that conditions (ii) and (iii) are satisfied. Hence, the types game is dominance solvable and the unique outcome that survives iterated elimination of strictly dominated strategies is the unique Nash equilibrium $(\widehat{\tau}, \widehat{\tau})$.

To establish the sign of $\widehat{\tau}$, recall that

$$f^i\left(\tau^i, \tau^j\right) = \Pi^i\left(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)\right).$$

Then $\widehat{\tau}$ is interior iff it satisfies the equation $f_i^i(\widehat{\tau}, \widehat{\tau}) = 0$. Otherwise, $\widehat{\tau} = \underline{\tau}$ $(\widehat{\tau} = \overline{\tau})$ if $f_i^i(\tau, \tau) < 0\ (> 0)$ for all $\tau \in (\underline{\tau}, \overline{\tau})$. Since $f_i^i(\cdot, \cdot)$ is continuous, these three cases completely characterize the possible values of $\widehat{\tau}$. Using this observation together with Assumptions B1 and B2, we conclude that $\widehat{\tau} > 0$ if $f_i^i(0,0) > 0$ and $\widehat{\tau} < 0$ if $f_i^i(0,0) < 0$. Thus to determine the sign of $\widehat{\tau}$, it suffices to determine the sign of $f_i^i(0,0)$. To this end, note that $\widehat{x}^i$ and $\widehat{x}^j$ are C^1 in a neighborhood of $(0,0)$ by Assumption A2 and the implicit function theorem, and

$$\begin{aligned} f_i^i(0,0) &= \Pi_i^i\left(\widehat{x}^i(0,0), \widehat{x}^j(0,0)\right)\frac{\partial \widehat{x}^i}{\partial \tau^i}(0,0) + \Pi_j^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))\frac{\partial \widehat{x}^j}{\partial \tau^i}(0,0) \\ &= \Pi_j^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))\frac{\partial \widehat{x}^j}{\partial \tau^i}(0,0), \end{aligned}$$

where the second equality follows from the fact that $\Pi_i^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0)) = 0$.

[9] Note that under our assumptions, this is equivalent to the dominant diagonal condition in Assumption B2.

To find an expression for $\frac{\partial \widehat{x}^j}{\partial \tau^i}(0,0)$, recall that $\widehat{x}^i$ and $\widehat{x}^j$ solve the equations

$$\Pi_i^i\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j)\right)+B_i^i\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j),\tau^i\right)=0,$$
$$\Pi_j^j\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j)\right)+B_j^j\left(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j),\tau^j\right)=0.$$

In what follows, we will occasionally simplify by abusing notation slightly and denoting $\Pi_i^i(\tau^i,\tau^j)=\Pi_i^i(\widehat{x}^i(\tau^i,\tau^j),\widehat{x}^j(\tau^i,\tau^j))$ and analogously for B^i and other derivatives thereof. We will also make use of our normalization, $B^i(x^i,x^j,0)\equiv B^j(x^i,x^j,0)\equiv 0$, from which it follows that $B_{ii}^i(x^i,x^j,0)=B_{ij}^i(x^i,x^j,0)=B_{jj}^j(x^i,x^j,0)=B_{ji}^j(x^i,x^j,0)=0$ for any (x^i,x^j).

Now by the implicit function theorem,

$$\begin{aligned}&\frac{\partial \widehat{x}^j}{\partial \tau^i}(0,0)\\&=\frac{\left(\Pi_{ji}^j(\widehat{x}^i(0,0),\widehat{x}^j(0,0))+B_{ji}^j(\widehat{x}^i(0,0),\widehat{x}^j(0,0),0)\right)B_{i\tau^i}^i(\widehat{x}^i(0,0),\widehat{x}^j(0,0),0)}{J(0,0)}\\&=\frac{1}{J(0,0)}\Pi_{ji}^j(\widehat{x}^i(0,0),\widehat{x}^j(0,0),0)B_{i\tau^i}^i(\widehat{x}^i(0,0),\widehat{x}^j(0,0),0),\end{aligned}$$

where the second equality follows from our normalization, and

$$J(\tau^i,\tau^j)\equiv\det\begin{pmatrix}\Pi_{ii}^i(\tau^i,\tau^j)+B_{ii}^i(\tau^i,\tau^j) & \Pi_{ij}^i(\tau^i,\tau^j)+B_{ij}^i(\tau^i,\tau^j)\\ \Pi_{ji}^j(\tau^i,\tau^j)+B_{ji}^j(\tau^i,\tau^j) & \Pi_{jj}^j(\tau^i,\tau^j)+B_{jj}^j(\tau^i,\tau^j)\end{pmatrix}.$$

Again by our normalization,

$$J(0,0)=\det\begin{pmatrix}\Pi_{ii}^i(\widehat{x}^i(0,0),\widehat{x}^j(0,0)) & \Pi_{ij}^i(\widehat{x}^i(0,0),\widehat{x}^j(0,0))\\ \Pi_{ji}^j(\widehat{x}^i(0,0),\widehat{x}^j(0,0)) & \Pi_{jj}^j(\widehat{x}^i(0,0),\widehat{x}^j(0,0))\end{pmatrix}.$$

Substituting and using the symmetry yields

$$\begin{aligned}&f_i^i(0,0)\\&=\frac{1}{J(0,0)}\Pi_j^i\left(\widehat{x}^i(0,0),\widehat{x}^j(0,0)\right)\Pi_{ij}^i\left(\widehat{x}^i(0,0),\widehat{x}^j(0,0)\right)\\&\quad\times B_{i\tau^i}^i\left(\widehat{x}^i(0,0),\widehat{x}^j(0,0),0\right).\end{aligned}$$

The desired conclusion then follows by noting that $J(0,0)>0$ as a consequence of Assumptions A1 and A2. □

When there is a unique type in the support of the limiting distribution, the interpretation of the sign of this type is largely a matter of conventions chosen in the modeling of the disposition family B. Hence the crucial point is whether this value is zero, corresponding to a population of payoff-maximizing agents, or non-zero. The second part of Lemma 1 illustrates a clear intuition for determining the value of this type. For example, when agents' actions are strategic complements in the underlying game, a higher action by one agent leads to a higher action by the

other. When there are positive externalities in the underlying game, this benefits the agents, and when types and actions are complements in the disposition function, a higher type induces a higher action by the agent. Together these forces lead to the emergence of a positive type in the limit. Similarly, negative externalities, strategic substitutability of actions, or substitutes in the type-action interaction reverse this prediction.

2.4 The evolutionary dynamics of dispositions

To study how dispositions evolve, suppose that there is a continuum of individuals of different types. We consider a continuous-time model in which at each point $t \geq 0$ in time, this population is characterized by the distribution $G_t \in \Delta(T)$, where $\Delta(T)$ denotes the set of Borel probability distributions over T. We assume that the initial distribution G_0 has full support over T. At each instant in time, individuals are randomly matched in pairs to play the game Γ. The average fitness levels of individuals of type τ at time t is then given by

$$\int f(\tau, \tau')dG_t(\tau').$$

Recall that, due to the symmetry in the types game, we have dropped the superscripts i and j here, and will do so in what follows.

We assume that the selection dynamics are monotonically increasing in average fitness. That is, we assume that the distribution of types evolves according to the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t(S) = \int_S g(\tau, G_t)dG_t(\tau), \quad S \subseteq T \text{ Borel measurable,} \tag{3}$$

where $g : T \times \Delta(T) \to \mathbf{R}$ is a continuous growth-rate function satisfying

$$g(\tau, G_t) > g(\tilde{\tau}, G_t) \iff \int f(\tau, \tau')dG_t(\tau') > \int f(\tilde{\tau}, \tau')dG_t(\tau'). \tag{4}$$

To ensure that G_t remains a probability measure for each t, we also assume that g satisfies

$$\int_T g(\tau, G_t)dG_t = 0 \quad \text{for each } t. \tag{5}$$

Equations (3)–(5) reflect the idea that the proportion of more successful types in the population increases from one period to another at the expense of less successful types. This may be due to the fact that more successful individuals have more descendants, who then inherit their parents' preferences either genetically or by education. An alternative explanation is that the decision rules of more successful individuals are imitated more often.

The same mathematical formulation is also compatible with the assumption that successful types translate into stronger *influence* rather than numerical proliferation. Under this interpretation, not all individuals are matched to play in each

period; instead more successful individuals take part in a larger share of the economic interactions, and so are matched to play with a higher probability.

To guarantee that the differential equation (3) has a well-defined solution, we require some additional regularity conditions on the selection dynamics as follows.

Definition 1 *(Regular dynamics) A growth rate function $g : T \times \Delta(T) \to \mathbf{R}$ is called regular if g can be extended to the domain $T \times Y$, where Y is the set of signed Borel measures with variational norm smaller than 2, such that on this extended domain, g is uniformly bounded and uniformly Lipschitz continuous. That is, there exist constants $M, K > 0$ such that*

$$\sup_{\tau \in T} |g(\tau, G_t)| < M, \quad \sup_{\tau \in T} \left|g(\tau, G_t) - g(\tau, \widetilde{G}_t)\right| < K \left\|G_t - \widetilde{G}_t\right\|, \quad \forall G_t, \widetilde{G}_t \in Y,$$

where $\|\mu\| = \sup_{|h| \leq 1} \left|\int_{\mathbf{R}} h d\mu\right|$ denotes the variational norm of the signed measure μ.

The dynamics $\{G_t, t \geq 0\}$ are called regular if G_t solves the differential equation (3) for some regular growth rate function g.

Oechssler and Riedel (2001, Lemma 3) prove that regularity of the dynamics guarantees that the mapping $G \mapsto \int_T g(\cdot, G) dG$ is bounded and Lipschitz continuous in the variational norm, from which it follows that the differential equation in the space of distributions $\Delta(T)$ defined by

$$\frac{\mathrm{d}}{\mathrm{d}t} G_t(S) = \int_S g(\tau, G_t) dG_t(\tau), \quad S \subseteq T \text{ Borel measurable},$$

has a unique solution for any initial distribution G_0.

A special case of the class of growth rate functions that we consider is the familiar replicator dynamics introduced by Taylor and Jonker (1978) for distributions with a finite support, and by Oechssler and Riedel (2001) for general distributions. 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 our general class of selection dynamics, we now seek to answer the following question: starting from some initial distribution, G_0, how will the distribution of types G_t evolve over time under such regular, payoff-monotonic dynamics? To provide an answer, we first establish the following theorem, which generalizes results in Samuelson and Zhang (1992) to the case of games with infinitely many strategies. This theorem establishes the result that any serially dominated strategy must eventually become extinct under any payoff-monotonic selection dynamics. As a particular implication, in a dominance solvable game any such dynamics must converge in distribution, and converge to a point mass at the unique action surviving the iterated elimination of strictly dominated strategies.

To state this result we need some additional notation. In keeping with our basic framework and applications, we state the result for symmetric two-player games with a compact strategy space; note however that the result carries over to asymmetric games with virtually the same proof. We consider a symmetric game with common strategy space $T \subset \mathbf{R}$ and payoff function $f : T \times T \to \mathbf{R}$. Let D denote

the set of serially dominated strategies in this game, so that $D = \cup_{n=0}^{\infty} D_n$ where $D_0 = \emptyset$ and for $n \geq 1$,

$$D_n = \left\{ t \in T \setminus D_{n-1} : \exists s \in T \setminus D_{n-1} \text{ such that } f(s,r) > f(t,r) \ \forall r \in T \setminus D_{n-1} \right\}.$$

Analogously, let U denote the set of serially undominated strategies in this game, so that $U = T \setminus D$. Equivalently, $U = \cap_{n=0}^{\infty} U_n$ where $U_0 = T$ and for $n \geq 1$,

$$U_n = \left\{ t \in T \setminus D_{n-1} : \forall s \in T \setminus D_{n-1} \ \exists r \in T \setminus D_{n-1} \text{ s.t. } f(t,r) \geq f(s,r) \right\}.$$

Theorem 1 *Let $T \subset \mathbf{R}$ be a compact set of strategies, $f : T \times T \to \mathbf{R}$ be the continuous payoff function of a symmetric two-player game, and $g : T \times \Delta(T) \to \mathbf{R}$ a regular, payoff monotonic growth-rate function. Let G_t be the population dynamics defined by the differential equation*

$$\frac{d}{dt} G_t(S) = \int_S g(t, G_t) dG_t, \quad S \subseteq T \text{ Borel measurable}$$

given initial distribution G_0 with full support on T. For every strategy $d \in D$ there is a neighborhood $W_d \subset T$ such that $\lim_{t \to \infty} G_t(W_d) = 0$. In particular, if the game is dominance solvable, so that $U = \{u\}$ for some $u \in T$, then G_t converges in distribution to the unit mass at u.

Proof We prove by induction that for each n, U_n is compact, and for every strategy $d \in D_n$ there is a neighborhood $W_d \subset T$ 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 $D_1 = \emptyset$ as well, i.e., no strategies are strictly dominated, then the claim holds vacuously. So without loss of generality assume $D_1 \neq \emptyset$. Now suppose that the claim holds for $n < k$.

We first prove that U_k is compact. Since $U_k \subset T$ and T is compact, it suffices to show that U_k is closed. To that end, let $\{t_n\} \subset U_k$ and $t_n \to t$. Let $s \in T \setminus D_{k-1}$. Since $\{t_n\} \subset U_k$, for each n there exists $r_n \in U_{k-1}$ such that $f(t_n, r_n) \geq f(s, r_n)$. By the inductive hypothesis, U_{k-1} is compact, hence $\{r_n\}$ has a convergent subsequence. Without loss of generality, take $r_n \to r$ for some $r \in U_{k-1}$. Then since f is continuous, $f(t,r) \geq f(s,r)$. Hence $t \in U_k$, which shows that U_k is closed.

Next we show that for each $d \in D_k$ there is an open neighborhood W_d such that $\lim_{t \to \infty} G_t(W_d) = 0$. To this end, let $d \in D_k$ and let $x \in U_{k-1}$ be such that

$$f(x,y) - f(d,y) > 0 \quad \text{for all } y \in U_{k-1}.$$

Let

$$B = \left\{ y \in T | f(x,y) - f(d,y) \leq 0 \right\}.$$

Since f is continuous, B is a compact subset of T, and by choice of d and x, $B \subset D_{k-1}$. In particular, B is a *proper* subset of D_{k-1}, since D_{k-1} is open by the induction hypothesis. Now let

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

and

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

Since $B \subsetneq D_{k-1}$, it follows that $s > 0$.

By the induction assumption, for each $y \in C$ there exists a neighborhood W_y of y such that $G_t(W_y) \to 0$. Then $\{W_y : y \in C\}$ is an open cover of C. As C is compact, there is a finite subcover $\{W_{y_1}, \ldots, W_{y_K}\}$, and for each t, $G_t(C) \leq \sum G_t(W_{y_k})$. Thus $G_t(C) \to 0$ as $t \to \infty$.

Now note that by construction,

$$f(x, y) - f(d, y) > s \quad \text{for all } y \in D_{k-1} \setminus C$$

and

$$f(x, y) - f(d, y) > 0 \quad \text{for all } y \in T \setminus D_{k-1}.$$

Since f is continuous and U_{k-1} is compact, there exists $\bar{s} > 0$ and open neighborhoods $V_x \ni x$ and $W_d \ni d$ such that for every $x' \in \overline{V}_x$ and $d' \in \overline{W}_d$,

$$f(x', y) - f(d', y) \geq s/2 \quad \text{for all } y \in D_{k-1} \setminus C$$

and

$$f(x', y) - f(d', y) \geq \bar{s} \quad \text{for all } y \in T \setminus D_{k-1}.$$

Since f is continuous on the compact set T, there exists a bound M such that $|f(w, z)| \leq M$ for all $(w, z) \in T \times T$. Now set $\varepsilon = \min\left\{\frac{s}{2}, \bar{s}, \frac{1}{2}\right\}$. There exists $\bar{t}$ such that for each $t \geq \bar{t}$, $G_t(C) \leq \epsilon/8M$ and $G_t(T \setminus C) > 1 - \epsilon$. Then for any $x' \in \overline{V}_x$, $d' \in \overline{W}_d$ and $t \geq \bar{t}$,

$$\begin{aligned} f(x', G_t) - f(d', G_t) &= \int_T \big[f(x', y) - f(d', y)\big] dG_t \\ &= \int_C \big[f(x', y) - f(d', y)\big] dG_t + \int_{T \setminus C} \big[f(x', y) - f(d', y)\big] dG_t \qquad (6) \\ &> (-2M)\frac{\varepsilon}{8M} + \varepsilon(1 - \varepsilon) \geq -\frac{\varepsilon}{4} + \varepsilon\left(1 - \frac{1}{2}\right) = \frac{\varepsilon}{4}. \end{aligned}$$

By the continuity of f, (6) holds also when G_t is replaced by any probability measure $\mu \in A \equiv \overline{\{G_t\}_{t \geq \bar{t}}}$, the closure of $\{G_t\}_{t \geq \bar{t}}$ in the weak topology.

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 exists $\delta > 0$ such that

$$g(x', G_t) - g(d', G_t) \geq \delta \quad \text{for any } x' \in \overline{V}_x, \quad d' \in \overline{W}_d, \quad \text{and} \quad t \geq \bar{t}. \qquad (7)$$

Then (7) also holds 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 $t \geq \bar{t}$

$$\frac{\int_{\overline{V}_x} g(y, G_t)dG_t}{G_t(\overline{V}_x)} - \frac{\int_{\overline{W}_d} g(y, G_t)dG_t}{G_t(\overline{W}_d)} \geq \delta.$$

Hence, by (3), for $t \geq \bar{t}$,

$$\frac{G_t\left(\overline{V}_x\right)}{G_t\left(\overline{W}_d\right)} \geq \frac{G_{\bar{t}}\left(\overline{V}_x\right)}{G_{\bar{t}}(\overline{W}_d)} \exp[\delta(t - \bar{t})] \to \infty \text{ as } t \to \infty.$$

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

Although we will mainly focus on applications satisfying the general conditions for dominance solvability outlined in the previous section, Theorem 1 has a wide variety of implications beyond this particular setting. By Theorem 1, any serially dominated strategy will eventually become extinct under regular payoff-monotonic selection dynamics, so the support of any limiting distribution must be a subset of the set U of serially undominated strategies. In games where this set can be characterized or computed easily, Theorem 1 gives useful predictions regarding the dynamic evolutionary outcomes. Our main results give two different instances of this idea. The first focuses on settings where the types game is dominance solvable because it satisfies the standard concavity and dominant diagonal conditions in Assumptions A1–A2 and B1–B2.

Theorem 2 *Suppose that Assumptions A and B are satisfied. Then there exists a unique type $\widehat{\tau} \in T$ such that given any initial distribution of types with full support T, the distribution of types converges in distribution to a unit mass at $\widehat{\tau}$ under any regular, payoff-monotonic selection dynamics.*

Moreover, if $(\widehat{x}^i(0,0), \widehat{x}^j(0,0)) \in int\,(X^i \times X^j)$, then under Assumptions A1-A2 and B1-B2,

$$sign\,\widehat{\tau} = sign\, f_i^i(0,0)$$
$$= sign\, \Pi_j^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))\Pi_{ij}^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0))B_{i\tau^i}^i(\widehat{x}^i(0,0), \widehat{x}^j(0,0), 0).$$

Proof The first statement follows immediately from Theorem 1. The second follows from Lemma 1 and Theorem 1. □

A second broad class of games to which these results apply is games with strategic complementarities.[10] In these games, the set of serially undominated strategies is an interval, with endpoints given by the smallest and largest pure strategy Nash equilibria in the game. These endpoints are easily computed by standard best-response iteration algorithms. By Theorem 1, any types outside this set must asymptotically become extinct under any payoff-monotonic selection dynamics. In addition, under strategic complementarities any symmetric game with a unique symmetric equilibrium must be dominance solvable; Theorem 1 then implies that a unique type will survive under any payoff-monotonic selection dynamics. We collect these results below.

[10] Koçkesen et al. (2000a, b) study the effects of strategic complementarities in the evolutionary stability of interdependent preferences.

Theorem 3 *Suppose $T \subset \mathbf{R}$ is compact and $f : T \times T \to \mathbf{R}$ satisfies the single crossing property.*[11] *Let f be the payoff function of a symmetric two-player game, and $\{G_t, t \geq 0\}$ be regular payoff-monotonic dynamics such that G_0 has full support T. Then there exist $\underline{\tau}, \overline{\tau} \in T$ with $\underline{\tau} \leq \overline{\tau}$ such that $U = [\underline{\tau}, \overline{\tau}]$. For every $\tau \in T \setminus [\underline{\tau}, \overline{\tau}]$, there exists a neighborhood W_τ of τ for which $G_t(W_\tau) \to 0$ as $t \to \infty$. If there is a unique symmetric equilibrium $(\widehat{\tau}, \widehat{\tau})$ in this game, then G_t converges in distribution to the unit mass at $\widehat{\tau}$.*

Proof This follows from Theorem 1 and Theorem 12 of Milgrom and Shannon (1994). □

3 Applications

In this section we consider two applications of our main result. In the first application we interpret the dispositions as perception biases–players can differ from one another in terms of their perceptions of the effects of their actions. Depending on the context, this could represent over- or under-confidence, optimism or pessimism, or departures from rational expectations and Bayesian updating with a common prior. In our second application, we consider interdependent preferences, in which players differ with respect to the weight that they assign to their opponents' payoffs in their utility functions.[12]

3.1 The evolution of perception biases

Our first application can be viewed as an evolutionary explanation for the large body of evidence mentioned in the Introduction on the prevalence of optimism and overconfidence.[13] This application shows that over time, individuals who are unrealistically optimistic about the influence of their actions will grow in number at the expense of other types, including those with accurate perceptions, and will eventually take over the entire population.

To study the evolution of perception biases regarding own actions, 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 and j, choose actions $x^i, x^j \in \mathbf{R}$. These actions can be thought of as the degree of effort or the level of investment the individuals put into the interaction.[14]

[11] A function $f : T \times T \to \mathbf{R}$ satisfies the single crossing property provided that for each $\tau_1' > \tau_1$, if $f(\tau_1', \tau_2) \geq (>) f(\tau_1, \tau_2)$ for some $\tau_2 \in T$ then $f(\tau_1', \tau_2') \geq (>) f(\tau_1, \tau_2')$ for each $\tau_2' > \tau_2$. See Milgrom and Shannon (1994).

[12] In another important class of examples, Heifetz and Segev (2004) use these results to study the dynamic emergence of "tough" behavior in a salient class of bargaining mechanisms under asymmetric information.

[13] For an alternative exploration of optimism and self-confidence based on dynamic inconsistency, see for example Benabou and Tirole (1999a, b) and Brocas and Carrillo (1999). Compte and Postlewaite (2004) give another justification for welfare-improving effects of perception biases, based on an interaction between beliefs and performance.

[14] 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 becomes more involved.

Suppose the individuals' actual payoff functions are given by

$$\Pi^i\left(x^i, x^j\right) = \left(\alpha - bx^j - x^i\right)x^i, \qquad \Pi^j\left(x^i, x^j\right) = \left(\alpha - bx^i - x^j\right)x^j, \tag{8}$$

where $\alpha > 0$ and $b \in (-1, 1)$. Note that $\Pi^i_j(x^i, x^j) = -bx^i$ and $\Pi^i_{ij}(x^i, x^j) = b$. In this example x^i and x^j are both positive in the relevant range, thus when $b > 0$ the individuals impose negative externalities on one another (i.e., the larger is j's action, the lower is i's payoff), and moreover, actions are strategic substitutes in the sense of Bulow et al. (1985). In contrast, when $b < 0$, the individuals impose positive externalities on one another, and their actions are strategic complements.

Although the payoffs are symmetric, individuals may differ in the way they perceive the effects of their actions: pessimistic types underestimate the value of α, optimistic types overestimate it, and only realistic types assess it correctly. Specifically, individual i perceives the value of α to be

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

and analogously for j. We assume that $-\alpha \leq \underline{\tau} < 0 < \frac{\alpha}{5} < \overline{\tau}$ and $\frac{b^2\alpha}{4+2b-b^2} \in T$. We assume that $-\alpha < \underline{\tau}$ in order to ensure that α^i and α^j are both positive, and assume that $\frac{\alpha}{5} < \overline{\tau}$ in order to ensure that equilibria are interior. One interpretation is that $\tau \neq 0$ due to some inherent psychological bias. Alternatively, α^i and α^j might represent random variables with true mean α, with nonzero values of τ^i and τ^j arising as a result of departures from a common prior, from rational expectations, or from Bayesian information processing.

For simplicity, we will refer to the type τ as the individual's perception bias. We will say that a player is an *optimist* if $\tau > 0$, a *pessimist* if $\tau < 0$, and a *realist* if $\tau = 0$. 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 and α^j for α in equation (8), the perceived payoffs of the individuals can be written as

$$\begin{aligned} U^i\left(x^i, x^j, \tau^i\right) &= \left(\alpha - bx^j - x^i\right)x^i + \tau^i x^i, \\ U^j\left(x^i, x^j, \tau^j\right) &= \left(\alpha - bx^i - x^j\right)x^j + \tau^j x^j. \end{aligned} \tag{9}$$

Here the players' dispositions are $B^i(x^i, x^j, \tau^i) = \tau^i x^i$ and $B^j(x^i, x^j, \tau^j) = \tau^j x^j$.

A straightforward calculation establishes that the unique Nash equilibrium of the game with these utility functions is $(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j))$, where

$$\begin{aligned} \widehat{x}^i(\tau^i, \tau^j) &= \frac{2(\alpha + \tau^i) - b(\alpha + \tau^j)}{4 - b^2} \\ \widehat{x}^j(\tau^i, \tau^j) &= \frac{2(\alpha + \tau^j) - b(\alpha + \tau^i)}{4 - b^2}. \end{aligned} \tag{10}$$

Substituting $\widehat{x}^i(\tau^i, \tau^j)$ and $\widehat{x}^j(\tau^i, \tau^j)$ into (8) yields the resulting fitness functions

$$\begin{aligned} f^i(\tau^i, \tau^j) &\equiv \Pi^i(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) \\ &= \frac{\big(2(\alpha+\tau^i)-b(\alpha+\tau^j)\big)\big(2\alpha-(2-b^2)\tau^i-b(\alpha+\tau^j)\big)}{(4-b^2)^2}, \\ f^j(\tau^i, \tau^j) &\equiv \Pi^j(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) \\ &= \frac{\big(2(\alpha+\tau^j)-b(\alpha+\tau^i)\big)\big(2\alpha-(2-b^2)\tau^j-b(\alpha+\tau^i)\big)}{(4-b^2)^2}. \end{aligned} \tag{11}$$

We can now use Theorem 2 to characterize the asymptotic distribution of types in the population.

Proposition 1 *Consider the game described above, and suppose that* $\frac{b^2\alpha}{4+2b-b^2} \in T$. *For any initial distribution of types with full support* T, *the distribution of types converges in distribution to a unit mass at* $\widehat{\tau} = \frac{b^2\alpha}{4+2b-b^2}$ *under any regular payoff-monotonic selection dynamics.*

Proof By Theorem 2, it suffices to verify that Assumptions A1–A2 and B1–B2 are satisfied. From (9), it is easy to see the Assumptions A1 and A2 are satisfied. From (11) it is easy to verify that Assumptions B1 and B2 are also satisfied. Hence the types game can be solved by iterated elimination of strictly dominated strategies. The unique outcome that survives this process is

$$\widehat{\tau} = \frac{b^2\alpha}{4+2b-b^2} > 0,$$

found by solving the equation $f_i^i(\widehat{\tau}, \widehat{\tau}) = 0$. □

Proposition 1 shows that the population will converge to a unit mass at $\frac{b^2\alpha}{4+2b-b^2}$. Notice that this type is strictly positive unless $b = 0$. That is, aside from the case where $b = 0$, in which there is no strategic interaction between the players, evolution gives rise to players who consistently overestimate the returns to their actions. All other types, including types who perceive the returns to their actions accurately, become extinct asymptotically under any regular payoff-monotonic selection dynamics.

The intuition underlying the evolution of optimistic types is fairly straightforward. Optimistic types play more aggressively than realists or pessimists, and choose larger actions as they exaggerate the impact of their actions on their payoffs. When $b > 0$, players' actions are strategic substitutes, so the aggressive behavior of optimists induces rivals to play soft. When $b > 0$ the players' actions have negative externalities on their opponents, so the soft behavior of his opponents benefits the optimistic player. Instead 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 opponents when $b < 0$, the aggressive behavior of rivals benefits the aggressive player in this case. Thus regardless of the sign of b, optimists gain a strategic advantage here. 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 moderate optimists, so on average, "cautiously" optimistic individuals fare better than individuals with other perceptions and therefore gradually take over the entire population.

Proposition 1 provides an evolutionary explanation for various 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 contradicts these views.[15] 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, or converge to a common prior.

In a similar but slightly more complicated setting, our results can be applied to give a fully dynamic version of the results of Kyle and Wang (1997) on the survivability of overconfidence in financial markets. In a duopoly version of the Kyle (1985) model, Kyle and Wang consider traders who have different noisy signals about the liquidation value of an asset, and relax the rational expectations assumption by allowing traders to have different distributions over signals. In the analogue of our types game that results, when each trader uses a linear pricing strategy, each trader's best response function has a slope less than one in absolute value, and the game is dominance solvable. Thus by our Theorem 2, any regular payoff-monotonic selection dynamics results in convergence to a unit mass at the unique type that is the dominance solution. This type corresponds to optimistic beliefs regarding the liquidation value of the asset in their model. As Kyle and Wang note, this has a number of interesting implications regarding the design of incentive contracts for managers and the desirability of delegated fund management; moreover, the fully dynamic results obtained using this methodology conform with some evidence indicating that more experienced traders may in fact be more overconfident than newcomers (see Griffen and Tversky 1992).

3.2 Interdependent preferences

The second application of our results is to the case where individuals care not only about their own material payoffs but also about the material payoffs of others. This could be due to factors like altruism or spite, or to concerns about relative rather than absolute payoffs.

To study the evolution of interdependent preferences, we consider the same setting as in section 3.1: a large population of individuals are continuously and randomly matched in pairs to interact with one another. In every pairwise interaction, the individuals i and j choose actions $x^i, x^j \in \mathbf{R}$ and their actual payoffs, $\Pi^i(x^i, x^j)$ and $\Pi^j(x^i, x^j)$, are given by (8), where now, $b \in (-1/2, 0) \cup (0, 2/5)$.

[15] 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).

Players' perceived payoffs will be a weighted average of their own payoff and their rival's payoff. The perceived payoffs of the individuals are given by

$$\begin{aligned} U^i\left(x^i, x^j, \tau^i\right) &= \Pi^i\left(x^i, x^j\right) + \tau^i \Pi^j\left(x^i, x^j\right) \\ U^j\left(x^i, x^j, \tau^j\right) &= \Pi^j\left(x^i, x^j\right) + \tau^j \Pi^i\left(x^i, x^j\right), \end{aligned} \tag{12}$$

where the parameter $\tau \in [-1, 1]$ is the player's type and measures the degree to which he cares about his opponent's material payoff. Here a player is altruistic when $\tau > 0$ and spiteful when $\tau < 0$.

Bester and Güth (1998), Bolle (2000), and Possajennikov (2000) have used the same setting to study the evolutionary stability of interdependent preferences. Bester and Güth (1998) restrict attention to the cases in which $b \in (-1, 1)$ and $\tau^i, \tau^j \in [0, 1]$, and show that the unique evolutionarily stable outcome is such that $\tau^i = \tau^j = -\frac{b}{2+b}$ if $b < 0$ and $\tau^i = \tau^j = 0$ if $b > 0$. Bolle (2000) shows that if players are allowed to be spiteful, so that $\tau^i, \tau^j \in (-\infty, 1]$, then the unique evolutionarily stable outcome is $\tau^i = \tau^j = -\frac{b}{2+b}$. Possajenikov (2000) shows that this result extends to cases where $b \in [-2, 1)$ or $b > 2$. Using Theorem 2, we can extend these results by showing that if $b \in (-1/2, 0) \cup (0, 2/5)$, then $\tau^i = \tau^j = -\frac{b}{2+b}$ is not only the unique evolutionarily stable outcome, but also the outcome to which the population will converge in distribution starting from any arbitrary initial distribution of types with support $[-1, 1]$.[16]

To see this, note first that the unique Nash equilibrium in the game given (τ^i, τ^j) is $(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j))$, where

$$\begin{aligned} \widehat{x}^i\left(\tau^i, \tau^j\right) &= \frac{a\left(2 - b(1 - \tau^i)\right)}{4 - b^2(1 - \tau^i)(1 - \tau^j)} \\ \widehat{x}^j\left(\tau^i, \tau^j\right) &= \frac{a\left(2 - b(1 - \tau^j)\right)}{4 - b^2(1 - \tau^i)(1 - \tau^j)}. \end{aligned} \tag{13}$$

Substituting $\widehat{x}^i(\tau^i, \tau^j)$ and $\widehat{x}^j(\tau^i, \tau^j)$ into (12) yields the resulting fitness functions[17]

$$\begin{aligned} f^i(\tau^i, \tau^j) &\equiv \Pi^i(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) \\ &= \frac{a^2\left(2 - b(1 + \tau^i)\right)\left(2 - b(1 - \tau^i) - b^2\tau^i(1 + \tau^j)\right)}{\left(4 - b^2(1 + \tau^i)(1 + \tau^j)\right)^2}, \\ f^j(\tau^i, \tau^j) &\equiv \Pi^j(\widehat{x}^i(\tau^i, \tau^j), \widehat{x}^j(\tau^i, \tau^j)) \\ &= \frac{a^2\left(2 - b(1 + \tau^j)\right)\left(2 - b(1 - \tau^j) - b^2\tau^j(1 + \tau^i)\right)}{\left(4 - b^2(1 + \tau^j)(1 + \tau^i)\right)^2}. \end{aligned} \tag{14}$$

[16] We need to restrict b and assume that $b \neq 0$ to ensure that the type game is dominance solvable, and hence that we can uniquely characterize the support of the limiting distribution. Bolle (2000) and Possajennikov (2000) were able to consider a larger range of values of b because instead they use the static notion of evolutionary stability.

[17] These functions coincide with equation (4) in Possajennikov (2000).

We can now use our general results to characterize the limiting distribution of types in this setting.

Proposition 2 *Consider the game described above with* $b \in (-1/2, 0) \cup (0, 2/5)$ *and* $T = [-1, 1]$. *For any initial distribution of types with full support* T, *the distribution of types converges in distribution to a unit mass at* $-\frac{b}{2+b}$ *under any regular payoff-monotonic selection dynamics.*

Proof By Theorem 2, it suffices to verify that Assumptions A1–A2 and B1–B2 are satisfied. From (12), it is easy to see the Assumptions A1 and A2 are satisfied. To check Assumption B1, note that straightforward differentiation of (14) yields

$$\frac{\partial^2 f^i}{\left(\partial \tau^i\right)^2}(\tau^i, \tau^j) = -\frac{2a^2b^2\left(2 - b(1+\tau^j)\right)}{\left(4 - b^2(1+\tau^i)(1+\tau^j)\right)^4} M\left(\tau^i\right),$$

where

$$M\left(\tau^i\right) \equiv 8 - 4b\left(1 + b(1-\tau^i)\right)(1+\tau^j) + b^3\left(4 - b - (2+b)\tau^i\right)(1+\tau^j)^2.$$

To determine the sign of $\frac{\partial^2 f^i}{\left(\partial \tau^i\right)^2}(\tau^i, \tau^j)$, note that

$$\frac{\partial M}{\partial \tau^i}(\tau^i) = b^2(1+\tau^j)\left(4 - b(2+b)(1+\tau^j)\right) \geq 0,$$

where the inequality follows because $\tau^j \geq -1$, $b < 1/2$, and $\tau^j \leq 1$. Hence for any $\tau^i \in T$,

$$M\left(\tau^i\right) \geq M(-1) = 8 - 4b\,(1+2b)\left(1+\tau^j\right) + 6b^3\left(1+\tau^j\right)^2.$$

Now suppose that $b > 0$. Then

$$M\left(\tau^i\right) \geq M(-1) \geq 8 - 4b\,(1+2b)\left(1+\tau^j\right) > 0,$$

where the last inequality follows because $b < 1/2$ and $\tau^j \leq 1$. On the other hand, if $b < 0$, then

$$M(\tau^i) \geq M(-1) \geq 8 + 6b^3(1+\tau^j)^2 \geq 5,$$

where the last inequality follows because $b < 1/2$ and $\tau^j \leq 1$. Hence $M(\tau^i) > 0$ for all $\tau^i \in [-1, 1]$, implying that $\frac{\partial^2 f^i}{\left(\partial \tau^i\right)^2}(\tau^i, \tau^j) < 0$. Hence Assumption B1 holds.

Since f^i is concave in τ^i, the best-response function of player i in the types game is

$$BR^i\left(\tau^j\right) \equiv -\frac{b\,(2-b)\,(1+\tau^j)}{4 - b(2+b)(1+\tau^j)}.$$

The best-response function of player j is analogous. To prove that Assumption B2 holds, it suffices to show that the slope of $BR^i(\cdot)$ is less than 1 in absolute value. To this end, note that

$$\frac{dBR^i}{d\tau^j}\left(\tau^j\right) = \frac{-4b\,(2-b)}{\left(4-b(2+b)(1+\tau^j)\right)^2}.$$

This expression is decreasing in τ^j, and hence is maximized at $\tau^j = -1$ and minimized at $\tau^j = 1$. For $b < 0$ this slope is positive, so the maximum absolute value occurs at $\tau^j = -1$, where the value is $\frac{-b(2-b)}{4} < 1$ since $b > -\frac{1}{2}$. For $b > 0$ this slope is negative, so the maximum absolute value occurs at $\tau^j = 1$, where the absolute value is

$$\frac{b(2-b)}{[2-b(2+b)]^2}.$$

It is straightforward to verify that this is less than 1 for $b < \frac{2}{5}$.

Hence by Theorem 2 the types game is dominance solvable. The unique outcome that survives this process is

$$\widehat{\tau} \equiv -\frac{b}{2+b},$$

found by solving the equation $f_i^i(\widehat{\tau}, \widehat{\tau}) = 0$. □

Proposition 2 shows that if $b < 0$ (i.e., agents impose positive externalities on one another), the population will converge over time to a unit mass at some moderate level of altruism, as $\widehat{\tau} > 0$. This level of altruism will be greater the greater are the positive externalities that the agents impose on one another. On the other hand, if $b > 0$ (i.e., agents impose negative externalities on one another), the population will converge to a moderate level of spite, as $\widehat{\tau} < 0$.[18]

4 Imperfect observability

Thus far, we have assumed that players in the basic game play a Nash equilibrium given their perceived payoff functions. One justification for this assumption is that players' perceived payoffs are perfectly observed. Of course, by standard arguments, Nash equilibrium play does not necessarily require observability of payoffs. If the interaction lasts several rounds, play can converge to a Nash equilibrium even if players have very limited knowledge or adapt their behavior myopically. This may be the case, for example, if the players follow some version of fictitious play (see e.g., Fudenberg and Levine 1998). In particular, in the setting of our main result, naive best-response dynamics yields play that converges to the unique Nash equilibrium for a given pair of types. Thus provided evolution occurs on a slower

[18] Note that this result illustrates the larger point of Heifetz et al. (et al.) regarding the generic emergence of dispositions. Here we are able to show that various different types of dispositions — in this case both perception biases and interdependent preferences — may emerge in the same basic setting of material payoffs given by this simple duopoly model.

time scale than learning in a given population, the Nash assumption may not be wildly inaccurate.[19]

In this section, we pursue further the possibility that preferences may not be perfectly observed. We consider two different settings: (i) preferences are observed only in a fraction of encounters, and (ii) players may be engaged in costly signaling regarding their preferences. We develop general results analogous to Theorems 1 and 2 for models with partial observability, and illustrate their application in a version of the example of section 3. We also show how the general results of Theorem 1 can be used to study a version of this example that includes costly signaling of preferences.

4.1 Partial observability

In this subsection we consider a setting in which preferences are observed in some exogenously specified fraction $1-\rho$ of interactions, but are completely unobserved in the remaining fraction ρ of interactions, where $\rho \in (0, 1]$. When preferences are not observed, players are assumed to play a Bayesian equilibrium, with common knowledge of the distribution of types in the population. We derive a general result along the lines of Theorem 1, and then demonstrate how it can be applied to a version of the example of section 3.1 incorporating partial observability.

We consider a class of two-player symmetric Bayesian games characterized by a payoff function of the form $f : T \times T \times (0, 1) \times T \to \mathbf{R}$. For each fixed $\rho \in (0, 1]$ indexing the fraction of interactions in which payoffs are unobservable, $f(\tau, \tau', \rho, \omega)$ gives the payoff to a player of type τ when facing a player of type τ', given that the average type in the current population is ω.

As in section 2.3, we consider the evolution of the population under payoff-monotonic selection dynamics. At each point $t \geq 0$ in time, this population is characterized by the distribution $G_t \in \Delta(T)$, with initial distribution G_0 having full support over T. The average fitness levels of individuals of type τ at time t is then given by

$$\int f(\tau, \tau', \rho, \bar{\tau}_{G_t}) \mathrm{d}G_t(\tau'),$$

where $\bar{\tau}_{G_t} \equiv \int \tau' dG_t(\tau')$ is the average type in the population at time t.

We assume that the distribution of types evolves according to the differential equation (3) where $g : T \times \Delta(T) \to \mathbf{R}$ is a continuous growth-rate function satisfying

$$\begin{aligned} & g(\tau, G_t) > g(\tilde{\tau}, G_t) \\ & \iff \int f(\tau, \tau', \rho, \bar{\tau}_{G_t}) dG_t(\tau') > \int f(\tilde{\tau}, \tau', \rho, \bar{\tau}_{G_t}) dG_t(\tau'). \end{aligned} \tag{15}$$

We derive a general result that mirrors Theorem 1, but dominance solvability is more complicated in this setting due to the dependence of payoffs on the average type as well as the type of the opponent. Thus we focus on a set of sufficient

[19] For example, Al-Najjar et al. (2004) give an interesting foundation for the sunk cost fallacy in which learning and experimentation justify the Nash equilibrium assumption.

conditions analogous to Assumptions B1 and B2 that guarantee both a version of dominance solvability, and convergence to a unique type under any regular, payoff monotonic selection dynamics.

Assumption C For each i:

1. The best response $BR^i(\tau; \rho, \omega)$ is single-valued for each $(\tau, \rho, \omega) \in T \times (0, 1) \times T$
2. either
   (a) f^i has the single crossing property in (τ^i, τ^j) and in $(\tau^i, -\omega)$, or
   (b) f^i has the single crossing property in $(\tau^i, -\tau^j)$ and in (τ^i, ω)
3. BR^i is Lipschitz continuous in τ and ω with constants k_τ and k_ω satisfying $k_\tau + k_\omega < 1$

Taken together, these assumptions guarantee that for each fixed ρ and ω, the symmetric types game with payoff function $f(\cdot, \cdot, \rho, \omega)$ is dominance solvable. In addition, the strategic complementarities in Assumption C2 guarantee that the unique symmetric equilibrium in this game is monotone in ω, decreasing in the case of 2(a), or increasing in the case of 2(b). Together with the single crossing property of the payoff function, this monotonicity implies that we can find a decreasing sequence of intervals containing any strategies that are not serially dominated in the underlying Bayesian game. This sequence must converge to a point as a consequence of the Lipschitz continuity in Assumption C3.

Theorem 4 *Let $T \subset \mathbf{R}$ be compact, and let $f : T \times T \times (0, 1) \times T \to \mathbf{R}$ be the payoff function of a symmetric two-player Bayesian game satisfying Assumption C. Then there exists a unique type $\widehat{\tau} \in T$ such that given any initial distribution of types with full support T, the distribution of types converges in distribution to a unit mass at $\widehat{\tau}$ under any regular, payoff-monotonic selection dynamics.*

Proof Fix $\rho \in (0, 1)$, let $g : T \times \Delta(T) \to \mathbf{R}$ be a regular, payoff monotonic growth-rate function, and let G_t be the population dynamics defined by (15), with initial distribution G_0 having full support T. To prove that the population converges under G_t converges in distribution to a unit mass at $\widehat{\tau}$, let $\widehat{\tau}$ be defined implicitly by the equation

$$\widehat{\tau} = BR^i(\widehat{\tau}; \rho, \widehat{\tau}). \tag{16}$$

Assumption C guarantees that $\widehat{\tau}$ is well-defined.

The idea behind the rest of the proof is as follows. The types game played at each point in time depends on the current average type ω. Hence, we cannot prove the convergence result as in Theorem 1 and need to use a more involved argument. Yet, once it is determined that irrespective of the value of ω, types outside an interval $[\tau_\ell, \tau_h]$ are serially dominated and hence asymptotically become extinct, the average type ω will eventually converge to an interval $[\tau_\ell - \delta, \tau_h + \delta]$, where δ is some small positive number. [20] The fact that $\widehat{\tau} \in [\tau_\ell - \delta, \tau_h + \delta]$ enables us to show that further types are serially dominated and thus that types outside some

[20] A priori we cannot rule out the possibility that ω will either approach τ_ℓ from below or τ_h from above, and therefore always remain outside the interval $[\tau_\ell, \tau_h]$. As an intermediate step we first show instead that ω will converge to some larger interval $[\tau_\ell - \delta, \tau_h + \delta]$.

smaller interval $[\tau'_\ell, \tau'_h] \subset [\tau_\ell, \tau_h]$ also asymptotically become extinct. The crux of the argument is in showing that it is impossible for this iterative process to stop with an interval of positive length.

We explore the evolution of the distribution of player i's types. The evolution of player j's types is analogous. Let

$$\overleftarrow{\tau} = \inf\left\{\tau' > \widehat{\tau} : \forall \tau > \tau' \; \exists V_\tau \ni \tau, V_\tau \text{ open, s.t. } \lim_{t\to\infty} G_t(V_\tau) = 0\right\}, \quad (17a)$$

$$\underrightarrow{\tau} = \sup\left\{\tau' < \widehat{\tau} : \forall \tau < \tau' \; \exists V_\tau \ni \tau, V_\tau \text{ open, s.t. } \lim_{t\to\infty} G_t(V_\tau) = 0\right\}. \quad (17b)$$

Here we use the convention that $\overleftarrow{\tau} = \overline{\tau}$ if the infimum in (17a) ranges over an empty set, and similarly that $\underrightarrow{\tau} = \underline{\tau}$ if the supremum in (17b) ranges over an empty set.

Let $\varepsilon > 0$, and set $A = [\underline{\tau}, \underrightarrow{\tau} - \frac{\varepsilon}{2}] \cup [\overleftarrow{\tau} + \frac{\varepsilon}{2}, \overline{\tau}]$. Then A is a compact subset of T. For each $\tau \in A$, let V_τ be a neighborhood of τ as given in (17a, b). Then $\{V_\tau : \tau \in A\}$ is an open cover of A. Take a finite sub-cover $V_{\tau_1}, \ldots, V_{\tau_n}$. Since $\lim_{t\to\infty} G_t(V_{\tau_k}) = 0$ for each $k = 1, \ldots, n$, there exists a time t_ε such that for $t > t_\varepsilon$, $G_t(V_{\tau_k}) < \frac{\varepsilon}{2nM}$ for each $k = 1, \ldots, n$, where $M = \max\{\varepsilon, \overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right), (\underrightarrow{\tau} - \frac{\varepsilon}{2}) - \underline{\tau}\}$. Hence, for $t > t_\varepsilon$ we conclude

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

Therefore, for $t > t_\varepsilon$ the average type in the population, ω, satisfies the following inequalities:

$$\begin{aligned}\omega &< \frac{\varepsilon}{2M}\overline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) = \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2M}\left(\overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right)\right) \\ &\le \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} = \overleftarrow{\tau} + \varepsilon, \end{aligned} \quad (18)$$

and

$$\begin{aligned}\omega &> \frac{\varepsilon}{2M}\underline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\underrightarrow{\tau} - \frac{\varepsilon}{2}\right) = \left(\underrightarrow{\tau} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2M}\left(\left(\underrightarrow{\tau} - \frac{\varepsilon}{2}\right) - \underline{\tau}\right) \\ &\ge \left(\underrightarrow{\tau} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2} = \underrightarrow{\tau} - \varepsilon. \end{aligned} \quad (19)$$

These inequalities imply that for every $\varepsilon > 0$, there exists a time t_ε such that for every $t > t_\varepsilon$, $\omega \in [\underrightarrow{\tau} - \varepsilon, \overleftarrow{\tau} + \varepsilon]$.

Next, consider Assumption C2(a) and C2(b) in turn. In either case, Assumptions C2 and C3 imply that for each fixed value of ω, there exists a unique symmetric Nash equilibrium in the fitness game. Under Assumption C2(a), $BR^i(\cdot; \rho, \omega)$ is decreasing in ω . Hence, the "highest" symmetric Nash equilibrium in the fitness game is attained when $\omega = \max\{\underrightarrow{\tau} - \varepsilon, \underline{\tau}\}$ and the "lowest" equilibrium is attained when $\omega = \min\{\overleftarrow{\tau} + \varepsilon, \bar{\tau}\}$. Let the highest and lowest symmetric Nash equilibria be $(\overrightarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon})$ and $(\overrightarrow{\tau_\varepsilon}, \overrightarrow{\tau_\varepsilon})$, respectively. That is, $\overleftarrow{\tau_\varepsilon}$ and $\overrightarrow{\tau_\varepsilon}$ are the solutions to the equations $\overleftarrow{\tau_\varepsilon} = BR^i(\overleftarrow{\tau_\varepsilon}; \rho, (\underrightarrow{\tau} - \varepsilon) \vee \underline{\tau})$ and $\underrightarrow{\tau_\varepsilon} = BR^i(\underrightarrow{\tau_\varepsilon}; \rho, (\overleftarrow{\tau} + \varepsilon) \wedge \bar{\tau})$.

By Assumption C2(a), for $\omega < \widetilde{\omega}$ and $\tau^i < \widetilde{\tau}$,

$$f^i\left(\widetilde{\tau}, \tau^j; \rho, \omega\right) < f^i\left(\tau^i, \tau^j; \rho, \omega\right) \text{ implies } f^i(\widetilde{\tau}, \tau^j; \rho, \widetilde{\omega}) < f^i(\tau^i, \tau^j; \rho, \widetilde{\omega}),$$

and similarly

$$f^i(\widetilde{\tau}, \tau^j; \rho, \widetilde{\omega}) > f^i(\tau^i, \tau^j; \rho, \widetilde{\omega}) \text{ implies } f^i(\widetilde{\tau}, \tau^j; \rho, \omega) > f^i(\tau^i, \tau^j; \rho, \omega).$$

These inequalities imply in turn that types above $\overleftarrow{\tau_\varepsilon}$ are serially dominated for $t > t_\varepsilon$, while types below $\underrightarrow{\tau_\varepsilon}$ are serially dominated for $t > t_\varepsilon$. By Theorem 1, this implies that types outside $[\underrightarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon}]$ asymptotically become extinct. By the definition of $\overleftarrow{\tau}$ and $\underrightarrow{\tau}$, it follows 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, letting $\varepsilon \to 0$ yields

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

Subtracting the second inequality from the first yields:

$$\begin{aligned}
0 &\leq \overleftarrow{\tau} - \underrightarrow{\tau} \\
&\leq \overleftarrow{\tau_0} - \underrightarrow{\tau_0} \\
&= |\overleftarrow{\tau_0} - \underrightarrow{\tau_0}| \\
&= |BR^i(\overleftarrow{\tau_0}; \rho, \underrightarrow{\tau}) - BR^i(\underrightarrow{\tau_0}; \rho, \overleftarrow{\tau})| \\
&= |BR^i(\overleftarrow{\tau_0}; \rho, \underrightarrow{\tau}) - BR^i(\overleftarrow{\tau_0}; \rho, \overleftarrow{\tau}) + BR^i(\overleftarrow{\tau_0}; \rho, \overleftarrow{\tau}) - BR^i(\underrightarrow{\tau_0}; \rho, \overleftarrow{\tau})| \\
&\leq |BR^i(\overleftarrow{\tau_0}; \rho, \underrightarrow{\tau}) - BR^i(\overleftarrow{\tau_0}; \rho, \overleftarrow{\tau})| + |BR^i(\overleftarrow{\tau_0}; \rho, \overleftarrow{\tau}) - BR^i(\underrightarrow{\tau_0}; \rho, \overleftarrow{\tau})| \\
&\leq k_\omega|\overleftarrow{\tau} - \underrightarrow{\tau}| + k_\tau|\overleftarrow{\tau_0} - \underrightarrow{\tau_0}| \\
&\leq (k_\omega + k_\tau)|\overleftarrow{\tau_0} - \underrightarrow{\tau_0}|.
\end{aligned}$$

Since $k_\omega + k_\tau < 1$, $\overleftarrow{\tau_0} = \underrightarrow{\tau_0}$, and $\overleftarrow{\tau} = \underrightarrow{\tau} = \widehat{\tau}$ as desired.

When Assumption C2(b) holds the argument is similar. Now for $\omega < \widetilde{\omega}$ and $\tau^i < \widetilde{\tau}$,

$$f^i(\widetilde{\tau}, \tau^j; \rho, \omega) > f^i(\tau^i, \tau^j; \rho, \omega) \text{ implies } f^i(\widetilde{\tau}, \tau^j; \rho, \widetilde{\omega}) > f^i(\tau^i, \tau^j; \rho, \widetilde{\omega}),$$

and similarly

$$f^i\left(\widetilde{\tau}, \tau^j; \rho, \widetilde{\omega}\right) < f^i\left(\tau^i, \tau^j; \rho, \widetilde{\omega}\right) \text{ implies } f^i\left(\widetilde{\tau}, \tau^j; \rho, \omega\right) < f^i\left(\tau^i, \tau^j; \rho, \omega\right).$$

Since $BR^i(\cdot; \rho, \omega)$ is decreasing in τ^j and increasing in ω, the highest best-response of i intersects the lowest best-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

for $t > t_\varepsilon$. By Theorem 1, types outside $[\underset{\rightarrow}{\tau}_\varepsilon, \overleftarrow{\tau}_\varepsilon]$ asymptotically become extinct. By the definition of $\overleftarrow{\tau}$ and $\underset{\rightarrow}{\tau}$, it follows that $\overleftarrow{\tau} \leq \overleftarrow{\tau}_\varepsilon$ and $\underset{\rightarrow}{\tau} \geq \underset{\rightarrow}{\tau}_\varepsilon$ for every $\varepsilon > 0$. Since $\overleftarrow{\tau}_\varepsilon$ and $\underset{\rightarrow}{\tau}_\varepsilon$ are continuous functions of ε, letting $\varepsilon \to 0$ yields

$$\overleftarrow{\tau} \leq \inf_{\varepsilon>0} \overleftarrow{\tau}_\varepsilon \equiv \overleftarrow{\tau}_0 = BR^i(\underset{\rightarrow}{\tau}_0; \rho, \overleftarrow{\tau}),$$
$$\underset{\rightarrow}{\tau} \geq \sup_{\varepsilon>0} \underset{\rightarrow}{\tau}_\varepsilon \equiv \underset{\rightarrow}{\tau}_0 = BR^i(\overleftarrow{\tau}_0; \rho, \underset{\rightarrow}{\tau}).$$

Then as above,

$$\begin{aligned} 0 \leq \overleftarrow{\tau} - \underset{\rightarrow}{\tau} \leq \overleftarrow{\tau}_0 - \underset{\rightarrow}{\tau}_0 &= \left| \overleftarrow{\tau}_0 - \underset{\rightarrow}{\tau}_0 \right| \\ &= \left| BR^i(\underset{\rightarrow}{\tau}_0; \rho, \overleftarrow{\tau}) - BR^i(\overleftarrow{\tau}_0; \rho, \underset{\rightarrow}{\tau}) \right| \\ &\leq (k_\omega + k_\tau) |\overleftarrow{\tau}_0 - \underset{\rightarrow}{\tau}_0|. \end{aligned}$$

Since $k_\omega + k_\tau < 1$, we again conclude that $\overleftarrow{\tau}_0 = \underset{\rightarrow}{\tau}_0$, and $\overleftarrow{\tau} = \underset{\rightarrow}{\tau} = \widehat{\tau}$ as desired. □

As an example of this result, we extend the perception bias example of section 3.1 to allow for partial observability of any degree $\rho \in (0, 1]$.

Proposition 3 *Consider the example of section 3.1 with partial observability and suppose that* $\frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3} \in T$. *If the initial distribution of types has full support, then under any regular payoff-monotonic selection dynamics the distribution of types converges in distribution to a unit mass at*

$$\widehat{\tau} = \frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3}.$$

Proof See the Appendix. □

Proposition 3 shows that the emerging type is monotonic in the probability $(1 - \rho)$ of observability. Moreover, the disposition becomes asymptotically extinct, that is, $\widehat{\tau} = 0$, only in the extreme cases where either $\rho = 1$ (preferences are never observed), or $b = 0$ (there is no strategic interaction between the players).[21] This example also illustrates the generic emergence of dispositions, since for any $b \in (-1, 1) \setminus \{0\}$ and any $\rho \in (0, 1)$, the unique type supported in the limit distribution is not zero.

[21] In different but analogous settings, Dekel et al. (2005), Ely and Yilankaya (2001), Ok and Vega Redondo (2001) and Güth and Peleg (2001) also consider what corresponds to our extreme case of $\rho = 1$, when preferences are completely unobservable, and show that payoff-maximization is evolutionarily stable.

4.2 Costly signaling of preferences

The benefit of having a disposition is the influence it exerts on opponents' equilibrium behavior, achieved at the cost of departures from actual payoff maximization. This leads to the following natural question. Can a player enjoy the best of all worlds – signal a disposition to rivals but choose actions that maximize his or her actual payoff? If signaling a disposition to others were merely cheap talk, then the signal would be ignored by the opponents, and hence the actual asymptotic behavior in the population would converge to a Nash equilibrium of the underlying game without dispositions. This is essentially the argument of Acemoglu and Yildiz (2001).

In practice, however, the appearance of individuals often does convey information about their dispositions. This information may be transmitted by body language or by past behavior in similar encounters. One possibility for why this is the case is that it may be costly to conceal dispositions. For instance, Frank (1987, 1988) argues vividly that some physical tendencies, like the blush that follows lying, may be the observable symptoms of emotional arousal that reflects the operation of *automatic* physical reactions and hence may be a credible signals about character. Although a mutant for whom these functions are not automatic and may be able to consciously control these functions may enjoy the benefits of lying without being caught, would also pay a fitness cost as a result of reacting more slowly to predators and enemies. This cost would be larger the less automatic these reactions are.

To model this idea, we suppose that players generate signals m^i and m^j regarding their true types τ^i and τ^j. The multiplicity of Bayesian equilibria associated with such a game at each point in time preempts a general analysis. Hence we consider an extension of the example from section 3.1. Suppose that $m^i, m^j \in M$, where $M \subset \mathbf{R}$ is a (large) compact interval that contains 0. To capture the idea that deception is costly, assume that these signals entail fitness costs $c(\tau^i - m^i)^2$ and $c(\tau^j - m^j)^2$, where $c > 0$. The signaling cost increases with the gap between the signal and the player's true type. We suppose that the signals m^i and m^j evolve in parallel with the players' types τ^i and τ^j according to some regular payoff-monotonic selection dynamics starting from some initial distribution with support in the rectangle $T \times M$. That is, the effective types of the players are now two-dimensional, consisting of both their disposition parameters τ^i and τ^j, and their signals m^i and m^j. When players i and j interact, they first observe each other's signals, update their beliefs about each other's preferences, and then play a Bayesian equilibrium given these updated beliefs. We can then characterize the limits of this evolutionary process as follows.

Proposition 4 *Consider the example from section 3.1 with costly signaling and suppose that* $\frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c)-b^3} \in T$, $\frac{2(1+4c)b^2\alpha}{2(4+2b-b^2)(1+4c)-b^3} \in M$, *and the initial distributions have full support. Then the joint distribution of types and signals will converge under any regular payoff-monotonic selection dynamics to a unit mass at*

$$\widehat{\tau} = \frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c)-b^3},$$
$$\widehat{m} = \frac{2\,(1+4c)\,b^2\alpha}{2(4+2b-b^2)(1+4c)-b^3}.$$

Proof See the Appendix. □

Note that unless $b = 0$, so that there is strategic interaction between the players, $\widehat{\tau} > 0$. This implies that in our model, the players will generically have dispositions. Also note that $\widehat{\tau}$ coincides with the type to which the population converges under full observability only when $c \to \infty$, that is, only when deception is infinitely costly. Otherwise, $\widehat{\tau}$ is smaller when signaling is costly. Intuitively, this is because the cost of deception must now be added to the costs of having a disposition. Furthermore, note that for all $0 < c < \infty$, $\widehat{m} > \widehat{\tau}$, implying that the population will converge over time to a type whose appearance will exaggerate the true disposition. That is, players will project they are overconfident to a larger extent than they actually are.

5 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 section we show that this argument continues to hold even if occasionally, individuals play against nature instead of being engaged in strategic interactions.

To capture this idea in a simple way, we will stay within the basic framework of section 3.1. We assume now that the parameter b is a binomial random variable, where $b = 0$ with probability ρ, and $b \in (-1, 1) \setminus \{0\}$ with probability $1 - \rho$. When $b = 0$ there is no strategic interaction, as the payoff to i is independent of j's action, so individuals can only lose by having a bias in this case.

The expected payoff of type τ^i when facing type τ^j is now given by:

$$f^i(\tau^i, \tau^j; \rho) = (1-\rho)\frac{\left(2(\alpha+\tau^i) - b(\alpha+\tau^j)\right)\left(2\alpha - (2-b^2)\tau^i - b(\alpha+\tau^j)\right)}{\left(4-b^2\right)^2} + \rho\frac{\left(\alpha+\tau^i\right)\left(\alpha-\tau^i\right)}{4}. \tag{20}$$

The expression multiplied by $1 - \rho$ is the payoff in (11), and the expression multiplied by ρ is the payoff in (11) evaluated at $b = 0$.

The best-response functions in the types game are given by

$$BR^i(\tau^j; \rho) = \frac{2(1-\rho)b^2\left(2\alpha - b(\alpha+\tau^j)\right)}{16 - 8b^2 + \rho b^4}, \tag{21}$$

and analogously for player j. As before, the best-response functions are downward sloping 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 types game is dominance solvable. The unique type that survives this process is

$$\widehat{\tau}(\rho) \equiv \frac{2(1-\rho)b^2\alpha}{2\left(4+2b-b^2\right) - \rho b^3}. \tag{22}$$

In particular, 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 one extreme, when $\rho = 1$ (with probability 1 there is no strategic interaction) $\widehat{\tau}(1) = 0$, so the population converges to realism. At the other extreme, when $\rho = 0$ (with probability 1 there is strategic interaction) $\widehat{\tau}(0) = \frac{b^2\alpha}{4+2b-b^2}$, so the situation is as in section 3.1. The main point, however, is that as long as there is a positive probability of strategic interaction between individuals (i.e., $\rho < 1$), the qualitative results from section 3.1 continue to hold.

We collect these observations in the final proposition.

Proposition 5 *Consider the example of section 3.1 and suppose that* $b = 0$ *with probability* ρ, $b \in (-1, 1) \setminus \{0\}$ *with probability* $1 - \rho$, *and*

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

Then for any initial distribution of types with full support T *and any payoff-monotonic selection dynamics, the distribution of types converges in distribution to a unit mass at* $\widehat{\tau}(\rho)$. *Moreover,* $\widehat{\tau}(\cdot)$ *is decreasing in* ρ, *with* $\widehat{\tau}(0) = \frac{b^2\alpha}{4+2b-b^2}$ *and* $\widehat{\tau}(1) = 0$.

6 Conclusion

Our results have illustrated how the pressures of explicit, dynamic evolutionary processes can select a unique preference profile when preferences are influenced by the outcomes of strategic interactions. The fully dynamic nature of the evolutionary analysis developed here expands the predictive power of evolutionary models of preference formation and strategic delegation.

This paper is part of a large and growing body of work that focuses on the meaning and foundations of rationality in economic models. Standard models often assume that decision makers have well-defined, stable and self-interested preferences, and act optimally given these preferences. In contrast, a large body of experimental work, combined with casual observation, suggests that individuals often behave in ways that are inconsistent with these assumptions. One approach to explaining such behavior is to attribute it to various bounds on the rationality of individuals, such as limited computational ability, limited memory, and so on. Instead our work has followed an alternative, more recent approach by exploring the evolutionary foundations of preference formation. This approach, by focusing on the evolution of preferences, shows that in a variety of contexts individuals can actually obtain higher payoffs if they strive to maximize some distorted form of their actual payoffs.

The value of dispositions in many different settings suggests that when contemplating the design of particular institutions, such as markets, auctions, or committees, it may be important to consider not only the equilibrium behavior of payoff-maximizing agents, but the equilibrium behavior of individuals whose behavior is biased by various dispositions as well. Moreover, this work suggests

that institutions themselves may influence the long-run preferences of participating agents. Preliminary analysis in this vein includes Bar-Gill and Fershtman (2001), Güth and Ockenfels (2001), Fershtman and Heifetz (2006), and Heifetz et al. (2006). Exploring the extent to which preferences may be in part an endogenous feature of the particular institutional framework seems to present promising avenues for future research.

Appendix

Following are the proofs of Propositions 3 and 4.

Proof of Proposition 3 When preferences are mutually observable, the equilibrium actions are specified in (10). When preferences are unobservable, we look for a Bayesian Nash equilibrium in which each player forms a belief about her 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 is given by:

$$U^i\left(x^i, \overline{x}; \tau^i\right) = \left(\alpha + \tau^i - b\overline{x} - x^i\right) x^i. \tag{23}$$

The problem of player j is analogous. The best-responses of players i and j against $\overline{x}$ are:

$$BR^i\left(\overline{x}, \tau^i\right) = \frac{\alpha + \tau^i - b\overline{x}}{2}, \quad BR^j\left(\overline{x}, \tau^j\right) = \frac{\alpha + \tau^j - b\overline{x}}{2}. \tag{24}$$

On the equilibrium path, the beliefs of the two players about $\overline{x}$ must be correct. Taking expectations on both sides of equation (24), using ω to denote the average type in the (current) population, and solving for $\overline{x}$ yields:

$$\overline{x} = \frac{\alpha + \omega}{2 + b}.$$

That is, when a player cannot observe the other player's preferences, the player (correctly) anticipates that given ω, the rival will play on average $\overline{x}$. Substituting for $\overline{x}$ in $BR^i(\overline{x}; \tau)$ and $BR^j(\overline{x}; \tau^j)$ yields equilibrium actions

$$\widehat{x}^i = \frac{\alpha + \tau^i - b\frac{\alpha+\omega}{2+b}}{2}, \quad \widehat{x}^j = \frac{\alpha + \tau^j - b\frac{\alpha+\omega}{2+b}}{2}.$$

Given $\widehat{x}^i$ and $\widehat{x}^j$, the resulting payoff of player i when the types are mutually unobserved is

$$\left(\alpha - b\frac{\alpha + \tau^j - b\frac{\alpha+\omega}{2+b}}{2} - \frac{\alpha + \tau^i - b\frac{\alpha+\omega}{2+b}}{2}\right)\left(\frac{\alpha + \tau^i - b\frac{\alpha+\omega}{2+b}}{2}\right). \tag{25}$$

With probability $1 - \rho$, preferences are observed and individual i's payoff is as in (11), whereas with probability ρ preferences are not observed and i's payoff is given by (23). Hence the expected fitness of player i when the player's type is τ^i,

the type of player j is τ^j, and the current average type of player j in the population is ω is given by

$$f^i(\tau^i, \tau^j; \rho, \omega) = (1-\rho)\frac{\left(2(\alpha+\tau^i) - b(\alpha+\tau^j)\right)\left(2\alpha - (2-b^2)\tau^i - b(\alpha+\tau^j)\right)}{\left(4-b^2\right)^2} + \rho\left(\alpha - b\frac{\alpha+\tau^j - b\frac{\alpha+\omega}{2+b}}{2} - \frac{\alpha+\tau^i - b\frac{\alpha+\omega}{2+b}}{2}\right)\left(\frac{\alpha+\tau^i - b\frac{\alpha+\omega}{2+b}}{2}\right). \quad (26)$$

The expected fitness of player j is analogous.

As in the proof of Proposition 1, consider the fitness game in which i and j choose their types, τ^i and τ^j, to maximize their fitness. The best-response function of player i in this game is

$$BR^i(\tau^j; \rho, \omega) = \frac{2\alpha b^2(2-b)(1-\rho)}{\left(16-8b^2+\rho b^4\right)} + \frac{b\rho(2-b)^2(2+b)^2}{2\left(16-8b^2+\rho b^4\right)}\omega - \frac{b\left(\rho b^4 + 4b^2 - 12\rho b^2 + 16\rho\right)}{2\left(16-8b^2+\rho b^4\right)}\tau^j. \quad (27)$$

The best response of player j, $BR^j(\tau^i; \rho, \omega)$, is analogous.

In what follows we use Theorem 4 to prove that the population converges over time to a stable monomorphic type. To this end, let $\widehat{\tau}$ be defined implicitly by the equation

$$\widehat{\tau} = BR^i(\widehat{\tau}; \rho, \widehat{\tau}).$$

Solving this equation yields

$$\widehat{\tau} = \frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3}.$$

Note that $\widehat{\tau} \in T$ by the assumption that $\frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3} \in T$.

Next, note from equation (26)

$$\frac{\partial^2 f^i}{\partial\tau^i\tau^j}(\tau^i, \tau^j, \rho, \omega) = (1-\rho)\left[\frac{b(2-b^2)}{(4-b^2)^2} + \frac{-2b}{(4-b^2)^2}\right] + \rho\frac{-b}{2} = (1-\rho)\frac{-b^3}{(4-b^2)^2} + \rho\frac{-b}{2}$$

and

$$\frac{\partial^2 f^i}{\partial\tau^i\omega}(\tau^i, \tau^j, \rho, \omega) = \rho\left[\frac{b^2}{4(2+b)} + \frac{b}{4(2+b)} + \frac{b}{4(2+b)}\right] = \frac{\rho(b^2+2b)}{4(2+b)} = \frac{\rho b}{4}$$

From this calculation we conclude that when $b < 0$, f^i satisfies Assumption C2(a), while when $b > 0$, Assumption C2(b) is satisfied instead.

In both cases, (27) shows that BR^i is Lipschitz in τ^j and in ω with constants

$$k_\tau = \frac{|b|\left(\rho b^4 + 4b^2 - 12\rho b^2 + 16\rho\right)}{2\left(16 - 8b^2 + \rho b^4\right)} = \frac{|b|\left(\rho(b^4 - 8b^2 + 16) + 4b^2(1-\rho)\right)}{2\left(16 - 8b^2 + \rho b^4\right)},$$

and

$$k_\omega = \frac{|b|\rho\,(2-b)^2\,(2+b)^2}{2\left(16 - 8b^2 + \rho b^4\right)}.$$

Notice that

$$\begin{aligned} k_\tau + k_\omega &= \frac{|b|\rho(2-b)^2(2+b)^2 + |b|\rho(2-b)^2(2+b)^2 + |b|4b^2(1-\rho)}{2\left(16 - 8b^2 + b^4\right) + 2\rho b^4 - 2b^4} \\ &= \frac{2|b|\rho(2-b)^2(2+b)^2 + |b|4b^2(1-\rho)}{2\left(16 - 8b^2 + b^4\right) + 2\rho b^4 - 2b^4} \end{aligned}$$

and

$$\begin{aligned} &2(1-|b|\rho)(16 - 8b^2 + b^4) - (1-\rho)(2b^4 + 4|b|^3) \\ &\quad > 2(1-|b|\rho)(16 - 8b^2) - (1-\rho)4|b|^3 \\ &\quad > 2(1-|b|\rho)(16 - 8b^2) - (1-\rho)4b^2 \\ &\quad > 2(1-|b|\rho)(16 - 12b^2) > 0. \end{aligned}$$

From this we conclude that $k_\tau + k_\omega < 1$ as desired.

The result now follows from Theorem 4. □

Proof of Proposition 4 The proposition follows from Theorem 1 once we show that $(\widehat{\tau}, \widehat{m})$ is the only combination of strategies that survives iterative elimination of strictly dominated strategies (τ, m) in the fitness game.

Player i with type τ^i and signal m^i chooses an action x^i to maximize the expected perceived payoff

$$(\alpha + \tau^i - b\chi^j(m^i, m^j) - x^i)x^i,$$

where the expectation is taken over the actions of player j with the signal m^j and $\chi^j(m^i, m^j)$ is the (current) average action of these players. The problem of player j is analogous.

The best-responses of players i and j against $\chi^j(m^i, m^j)$ and $\chi^i(m^i, m^j)$, are

$$x^i = \frac{\alpha + \tau^i - b\chi^j(m^i, m^j)}{2}, \quad x^j = \frac{\alpha + \tau^j - b\chi^i(m^i, m^j)}{2}. \tag{28}$$

Let $\tau^i(m^i)$ and $\tau^j(m^j)$, respectively, be the (current) average types of player i with signal m^i and player j with signal m^j. Taking expectations on both sides of the equations in (28) yields:

$$\begin{aligned} \chi^i(m^i, m^j) &= \frac{\alpha + \tau^i(m^i) - b\chi^j(m^i, m^j)}{2}, \\ \chi^j(m^i, m^j) &= \frac{\alpha + \tau^j(m^j) - b\chi^i(m^i, m^j)}{2}. \end{aligned}$$

Solving these two equations yields

$$\chi^i(m^i, m^j) = \frac{2\alpha + 2\tau^i(m^i) - \alpha b - b\tau^j(m^j)}{4 - b^2},$$
$$\chi^j(m^i, m^j) = \frac{2\alpha + 2\tau^j(m^j) - \alpha b - b\tau^i(m^i)}{4 - b^2}.$$

Substituting $\chi^i(m^i, m^j)$ and $\chi^j(m^i, m^j)$ in (28) reveals that the equilibrium actions of players i and j are given by

$$\widehat{x}_i = \frac{\alpha + \tau^i - b\frac{2\alpha + 2\tau^j(m^j) - \alpha b - b\tau^i(m^i)}{4-b^2}}{2},$$
$$\widehat{x}_j = \frac{\alpha + \tau^j - b\frac{2\alpha + 2\tau^i(m^i) - \alpha b - b\tau^j(m^j)}{4-b^2}}{2}.$$

The resulting (current) average fitness of player i of type τ^i and signal m^i when meeting player j with signal m^j is therefore

$$f^i\left(\left(\tau^i, m^i\right), m^j\right)$$
$$= \left(\alpha - b\frac{2\alpha + 2\tau^j(m^j) - \alpha b - b\tau^i(m^i)}{4 - b^2} - \frac{\alpha + \tau^i - b\frac{2\alpha + 2\tau^j(m^j) - \alpha b - b\tau^i(m^i)}{4-b^2}}{2}\right)$$
$$\times \frac{\alpha + \tau^i - b\frac{2\alpha + 2\tau^j(m^j) - \alpha b - b\tau^i(m^i)}{4-b^2}}{2} - c(m^i - \tau^i)^2. \tag{29}$$

The corresponding average fitness of player j is analogous. Maximizing $f^i\left(\left(\tau^i, m^i\right), m^j\right)$ with respect to τ^i and $f^j((\tau^j, m^j), m^i)$ with respect to τ^j implies that among all types of player i with the signal m^i and among all types of player j with the signal m^j, those with the highest average fitness are

$$\widehat{\tau}^i(m^i) = \frac{4c}{1 + 4c}m^i, \quad \widehat{\tau}^j(m^j) = \frac{4c}{1 + 4c}m^j.$$

Therefore, under regular payoff-monotonic selection dynamics, the combination $\left(\widehat{\tau}^i(m^i), m^i\right)$ will have the highest growth rate among all types of player i with signal m^i, and the combination and $\left(\widehat{\tau}^j(m^j), m^j\right)$ will have the highest growth rate among all types of player j with signal m^j. This implies in turn that

$$\lim_{t\to\infty} \tau^i\left(m^i\right) = \widehat{\tau}^i(m^i) = \frac{4c}{1 + 4c}m^i,$$
$$\lim_{t\to\infty} \tau^j\left(m^j\right) = \widehat{\tau}^j(m^j) = \frac{4c}{1 + 4c}m^j. \tag{30}$$

Taking the limit of the expressions in (29) as $t \to \infty$ and using (30), yields

$$f^i(m^i, m^j) \equiv \lim_{t\to\infty} f^i\left(\left(\widehat{\tau}^i(m^i), m^i\right), m^j\right)$$
$$= \left(\alpha - b\frac{2\alpha + 2\frac{4c}{1+4c}m^j - \alpha b - b\frac{4c}{1+4c}m^i}{4-b^2} - \frac{\alpha + \frac{4c}{1+4c}m^i - b\frac{2\alpha+2\frac{4c}{1+4c}m^j - \alpha b - b\frac{4c}{1+4c}m^i}{4-b^2}}{2}\right)$$
$$\times \frac{\alpha + \frac{4c}{1+4c}m^i - b\frac{2\alpha+2\frac{4c}{1+4c}m^j - \alpha b - b\frac{4c}{1+4c}m^i}{4-b^2}}{2} - c\left(m^i - \frac{4c}{1+4c}m^i\right)^2.$$

The corresponding expression for player j is analogous.

Now consider fitness game in which players i and j choose their signals m^i and m^j to maximize their respective fitness, $f^i(m^i, m^j)$ and $f^j(m^j, m^i)$. The best response function of player i in this game is given by

$$BR^i(m^j) = \frac{2b^2\alpha(2-b)(1+4c)}{\left(4-b^2\right)^2 + 32c(2-b^2)} - \frac{8cb^3}{\left(4-b^2\right)^2 + 32c(2-b^2)}m^j, \quad (31)$$

and analogously for player j.

The slope of these best-response functions is less than 1 in absolute value. Since the strategy sets in this signal game are one-dimensional compact intervals (recall that $m^i, m^j \in M$, where M is a one-dimensional compact interval), and the functions f^i and f^j are smooth and strictly concave in the players' own strategies, as in Lemma 1 it follows from Moulin (1984, Theorem 1) that the types game can be solved by iterative elimination of strictly dominated strategies. The unique signal that survives this process is

$$\widehat{m} = \frac{2(1+4c)b^2\alpha}{2(4+2b-b^2)(1+4c) - b^3},$$

which can be found by setting $BR^i(m^j) = m^j = \widehat{m}$ in equation (31). By Theorem 1, the distribution of signals converges to a unit mass at $\widehat{m}$. Using (30), the resulting types are

$$\widehat{\tau}^i = \widehat{\tau}^j = \frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c) - b^3},$$

as claimed. □

References

Acemoglu, D., Yildiz, M.: Evolution of perceptions and play. mimeo, MIT 2001
Alchian, A.: Uncertainty, evolution and economic theory. J Polit Econ **58**, 211–221 (1950)
Al-Najjar, N., Baliga, S., Besanko, D.: The sunk cost bias and managerial pricing policies, working paper, Northwestern University 2004
Babcock, L., Loewenstein, G.: Explaining bargaining impasse: the role of self-serving biases. J Econ Perspect **11**(1), 109–126 (1997)
Bar-Gill, O., Fershtman, C.: The limit of public policy: endogenous preferences, Foerder working paper 5-01, Tel Aviv University 2001

Benabou, R., Tirole, J.: Self-confidence: intrapersonal strategies, mimeo, Princeton University 1999a
Benabou, R., Tirole, J.: Self-confidence and social interactions, mimeo, Princeton University 1999b
Benos, A.V.: Aggressiveness and survival of overconfident traders. J Financ Mark **1**, 353–383 (1998)
Bergman, N., Bergman, Y.: Ecologies of Preferences with envy as an antidote to risk-aversion in Bargaining, mimeo, The Hebrew University of Jerusalem 2000
Bester, H., Güth, W.: Is altruism evolutionarily stable?. J Econo Behav Org **34**(2), 211–221 (1998)
Bolle, F.: Is altruism evolutionarily stable? and envy and malevolence? – remarks on Bester and Güth. J Econo Behav Org **42**(1), 131–133 (2000)
Brocas, I., Carrillo, J.: Entry mistakes, entrepreneurial boldness and optimism, CEPR D.P. 2213 1999
Bulow, J., Geanakoplos, J., Klemperer, P.: Multimarket oligopoly: strategic substitutes and complements. J Polit Econ **93**(3), 488–511 (1985)
Camerer, C.: Behavioral game theory, experiments in strategic interaction, Princeton: Princeton University Press 2003
Camerer, C., Lovallo, D.: Overconfidence and excess entry: an experimental approach. Am Econ Rev **89**(1), 306–318 (1999)
Christensen-Szalanski, J.J., Bushyhead, J.: Physicians' use of probabilistic information in a real clinical setting. J Exp Psychol Hum Percept Perform **7**, 928–935 (1981)
Compte, O., Postlewaite, A.: Confidence-enhanced performance. Am Econ Rev **94**(5), 1536–1557 (2004)
Cooper, A.C., Dunkelberg, W., Woo, C.: Entrepreneurs' perceived chances of success. J Bus Venturing **3**, 97–108 (1988)
De Bondt, W., Thaler, R.: Does the stock market overreact?. J Financ **40**, 793–807 (1985)
Dekel, E., Scotchmer, S.: On the evolution of attitudes toward risk in winner-take-all games. J Econo Theory **87**, 95–124 (1999)
Dekel, E., Ely, J., Yilankaya, O.: Evolution of preferences. mimeo, Northwestern University 2005
Dufwenberg, M., Güth, W.: Indirect evolution vs. strategic delegation: a comparison of two approaches to explaining economic institutions. Eur J Polit Econ **15**, 281–295 (1999)
Ely, J., Yilankaya, O.: Nash equilibrium and the evolution of preferences. J Econ Theory **97**(2), 255–272 (2001)
Fershtman, C., Heifetz, A.: Read my lips, watch for leaps: preference equilibrium and political instability. The Econ J **116**, 246–265 (2006)
Fershtman, C., Judd, K.: Incentive equilibrium in oligopoly. Am Econ Rev **77**(5), 927–940 (1987)
Fershtman, C., Weiss, Y.: Why do we care about what others think about us?. In: Ben Ner, A., Putterman, L. (eds) Economics, values and organization, Cambridge: Cambridge University Press 1997
Fershtman, C., Weiss, Y.: Social rewards, externalities and stable preferences. J Publ Econ **70**, 53–74 (1998)
Frank, R.H.: If homo economicus could choose his own utility function, would he choose one with a conscience? Am Econ Rev **77**(4), 593–604 (1987)
Frank, R.H.: Passions within reason – the strategic role of the emotions. New York: W.W. Norton & Company 1988
Friedman, M.: Essays in positive economics. Chicago: University of Chicago Press 1953
Froot, K., Frankel, J.: Forward discount bias: is it an exchange risk premium? Q J Econ **104**, 139–161 (1989)
Fudenberg, D., Levine, D.: The theory of learning in games. Cambridge: MIT Press 1998
Griffen, D., Tversky, A.: The weighing of evidence and the determinants of confidence. Cognitive Psychol **24**, 411–435 (1992)
Güth W., Peleg, B.: When will payoff maximization survive? An indirect evolutionary analysis. J Evol Econ **11**, 479–499 (2001)
Güth W., Yaari, M.: Explaining reciprocal behavior in simple strategic games: an evolutionary approach. In: Witt, U. (ed) Explaining forces and changes: approaches to evolutionary economics. Michigan: University of Michigan Press 1992
Guthrie, C., Rachlinski, J., Wistrich, A.: Inside the judicial mind: heuristics and biases. Cornell Law Rev **86**, 777–830 (2001)

Güth W., Ockenfels, A.: The coevolution of morality and legal institutions – an indirect evolutionary approach, mimeo 2001
Heifetz A., Segev, E.: The evolutionary role of toughness in bargaining. Games Econ Behav **49**, 117–134 (2004)
Heifetz, A., Segev, E., Talley, E.: Market design with endogenous preferences. Games Econ Behav (in press) 2006
Heifetz, A., Shannon, C. Spiegel, Y.: What to maximize if you must. J Econ Theory (previous version, IBER working paper, 2003) (in press) 2006
Huck S., Oechssler, J.: The indirect evolutionary approach to explaining fair allocations. Games Econ Behav **28**, 13–24 (1998)
Huck S., Kirchsteiger, G., Oechssler, J.: Learning to like what you have – explaining the endowment effect. The Econo J (in press) 2006
Katz M.: Game-playing agents: unobservable contracts as precommitments. Rand J Econ **22**, 307–328 (1991)
Kidd, J.: The utilization of subjective probabilities in production planning. Acta Psychol **34**, 338–347 (1970)
Koçkesen, L., Ok, E.A., Sethi, R.: Evolution of interdependent preferences in aggregative games. Games Econ Behav **31**, 303–310 (2000a)
Koçkesen, L., Ok, E.A., Sethi, R.: The strategic advantage of negatively interdependent preferences. J Econ Theory **92**, 274–299 (2000b)
Kyle, A.S.: Continuous auctions and insider trading. Econometrica **53**, 1315–1336 (1985)
Kyle A.S., Wang, A.: Speculation duopoly with agreement to disagree: can overconfidence survive the market test? The J Financ **LII**, 2073–2090 (1997)
Milgrom, P., Shannon, C.: Monotone comparative statics. Econometrica, **62**, 157–180 (1994)
Moulin, H.: Dominance solvability and cournot stability. Math Soc Sci **7**(1), 83–102 (1984)
Myers, D.: Psychology, Fifth Edition, New York: Worth Publishers Inc 1998
Oechssler, J., Riedel, F.: Evolutionary dynamics on infinite strategy spaces. Econ Theory **17**, 141–162 (2001)
Ok, E.A., Vega-Redondo, F.: On the evolution of individualistic preferences: an incomplete information scenario. J Econ Theory **97**, 231–254 (2001)
Oskamp, S.: Overconfidence in case study judgments. J Consult Psychol **29**, 261–265 (1965)
Possajennikov, A.: On the evolutionary stability of altruistic and spiteful preferences. J Econo Behav Org **42**(1), 125–129 (preliminary version appeared as CentER working paper 9956, Tilburg University). 2000
Robson, A.R.: A biological basis for expected and non-expected utility. J Econ Theory **68**, 397–424 (1996a)
Robson, A.R.: The evolution of attitudes to risk: lottery tickets and relative wealth. Games Econ Behav **14**, 190–207 (1996b)
Rogers, A.R.: Evolution of time preference by natural selection. Am Econ Rev **84**, 460–481 (1994)
Rosen, J.: Existence and uniqueness of equilibrium points for concave n-person games. Econometrica, **33**, 520–534 (1965)
Russo, J., Schoemaker, P.: Managing overconfidence. Sloan Manag Rev **33**, 74–76 (1992)
Samuelson, L., Zhang, J.: Evolutionary stability in asymmetric games. J Econ Theory, **57**, 363–391 (1992)
Sandholm, W.: Preference evolution, two-speed dynamics, and rapid social change. Rev Econ Dynam **4**, 637–679 (2001)
Schelling, T.: The strategy of conflict, Cambridge: Harvard University Press 1960
Taylor P., Jonker, L.: Evolutionary stable strategies and game dynamics. Math Biosci **40**, 145–156 (1978)
Vega-Redondo, F.: The evolution of walrasian behavior. Econometrica **65**, 375–384 (1997)
Waldman, M.: Systematic errors and the theory of natural selection. Am Econ Rev **84**(3), 482–497 (1994)
Wason, P. C.: The importance of cognitive illusions. The Behav Brain Sci **4**, 356 (1981)
Weibull, J.: Evolutionary game theory. Cambridge: MIT Press 1995